A general perspective on the Metropolis–Hastings kernel

Christophe Andrieu*, Anthony Lee* and Sam Livingstone†

January 1, 2021

*School of Mathematics, University of Bristol, U.K.
†Department of Statistical Science, University College London, U.K.

Abstract

Since its inception the Metropolis–Hastings kernel has been applied in sophisticated ways to address ever more challenging and diverse sampling problems. Its success stems from the flexibility brought by the fact that its verification and sampling implementation rests on a local "detailed balance" condition, as opposed to a global condition in the form of a typically intractable integral equation. While checking the local condition is routine in the simplest scenarios, this proves much more difficult for complicated applications involving auxiliary structures and variables. Our aim is to develop a framework making establishing correctness of complex Markov chain Monte Carlo kernels a purely mechanical or algebraic exercise, while making communication of ideas simpler and unambiguous by allowing a stronger focus on essential features — a choice of embedding distribution, an involution and occasionally an acceptance function — rather than the induced, boilerplate structure of the kernels that often tends to obscure what is important. This framework can also be used to validate kernels that do not satisfy detailed balance, i.e. which are not reversible, but a modified version thereof.

Contents

1 Introduction

Assume one is interested in sampling from a probability distribution π, defined on some probability space $(\mathsf{Z}, \mathscr{Z})$. A Markov chain Monte Carlo algorithm (MCMC) consists of simulating a realization of a time-homogeneous Markov chain $(Z_0, Z_1, Z_2 \ldots)$, of say kernel P, with the property that the distribution of Z_n becomes arbitrarily close to π as $n \to \infty$ irrespective of the distribution of Z_0. A property the kernel P, or its components in the case of mixtures or composition of kernels, must satisfy is to leave the distribution π invariant, that is π should be a fixed point of the Markov kernel. This is often referred to as a "global balance" condition in the physics literature and is most often not tractable to verify. Instead one can consider the stronger "detailed balance" condition, or reversibility, a more tractable property due to its local character which has led in particular to the celebrated Metropolis-Hastings (MH) kernel (Metropolis et al., 1953; Hastings, 1970), the cornerstone of MCMC simulations, and a multitude of successful variations. It is difficult to overstate the importance of detailed balance when discussing the widespread application of MH kernels: one can view such a kernel as being defined by a pair (π, Q), where Q is a proposal Markov kernel, and the algorithm requires only simulation according to Q and computing densities associated with π and Q. This ease of use has lead to MH algorithms being used in increasingly sophisticated contexts, leading to sometimes spectacular practical improvements but also increased complexity when establishing correctness (which we will take throughout to mean ensure that π is left invariant by P) and communicating their structure. The aim of this paper is to develop a simple and general framework to address these issues. In particular, the proposed framework defines an invariant MH kernel Π using a triple (μ, ϕ, a), where μ is the invariant distribution of Π, ϕ is an involution and a is an acceptance function, and retains similar ease-of-use properties to those described above: one is required only to be able to simulate from an appropriate conditional distribution of μ, calculate ϕ and ratios of densities involving μ and ϕ.

1.1 Contributions

We consider a framework, extending Tierney (1998), for defining a μ-reversible Markov kernel Π of the Metropolis-Hastings type, which only requires the specification of a triplet (μ, ϕ, a) where μ is a probability measure on some space $(\mathsf{E}, \mathscr{E})$, $\phi \colon \mathsf{E} \to \mathsf{E}$ an involution, and $a \colon \mathbb{R}_+ \to [0, 1]$ an acceptance function. As we shall see, this covers most scenarios of interest where sampling from π as above is of interest by letting π be a marginal of μ. More specifically for $\xi = (\xi_0, \xi_{-0}) \sim \mu$ such that $\xi_0 \sim \pi$, ξ_{-0} is a set of instrumental random variables involved in the design of MH kernels–often referred to as "proposals" for standard algorithm, but we refrain from using this reductive terminology. Then the involution ϕ is applied, defining $\xi' := \phi(\xi_0)$, and ξ_0' is the next state of the Markov chain with a probability entirely determined by the triplet (μ, ϕ, a), or the Markov chain remains at ξ_0. What is remarkable is that a correct algorithm is mathematically entirely determined by this triplet–in particular there is, again at a theoretical level, no need to determine an expression for the "acceptance ratio": it exists!

Practical implementation requires determining a tractable expression for the acceptance ratio which is, fundamentally, of a measure theoretic nature. Measure theoretic arguments are often overlooked in the literature and indeed do not need to be considered in detail in most simple scenarios. However this is not the case for more involved cases, where such issues can lead to excruciating and *ad hoc* contortions, and we have made an effort here not to ignore them. We hope to convince the reader that doing so is truly valuable and brings both generality and clarity to the arguments. The background required is minimal and we provide key results in the text: extensive knowledge of measure theory is not a prerequisite to read the manuscript.

As we shall see we focus primarily on the choice of (μ, ϕ) since the choice of a is, at least theoretically, independent of the choice of (μ, ϕ) and can be determined optimally thanks to the results of Peskun (1973) and Tierney (1998) in the reversible setup and Christophe Andrieu and Livingstone (2019)

for nonreversible extensions. We revisit numerous examples, some particularly simple for pedagogical purposes, but also dedicate full sections (Sections 5 and 7) to popular examples which, we know, have baffled more than one researcher before. This includes the No U-Turn Sampler (Hoffman and Gelman, 2014), the extra-chance algorithm (Sohl-Dickstein, Mudigonda, and DeWeese, 2014; Campos and J. Sanz-Serna, 2015) or event chain algorithms (Michel, 2016). In fact, We provide generalizations and in some cases completely novel versions of these algorithms.

We neither address the issues of convergence to equilibrium or ergodic averages, nor answer the question of what is the best possible involution. These are completely separate issues but we note that the ideas of Thin, Kotelevskii, et al. (2020), or more generally adaptive MCMC (Christophe Andrieu and Thoms, 2008), could be used for the latter purpose while Durmus, Moulines, and Saksman (2017) and Thin, Durmus, et al. (2020) provide some ideas concerning general results to establish irreducibility and aperiodicity, the additional sufficient ingredients needed to ensure convergence. There are in our view too many degrees of freedom involved in the choice of good involutions, auxiliary variables and their distributions and we do not believe that a theorem can, yet, replace intuition, creativity and commonsense when designing good MCMC schemes. Our aim here is rather to make checking that one's intuition is correct a purely algebraic exercise, removing in particular the need to revisit common points every time the question of correctness arises, while helping with efficient and unambiguous communication of potentially very complex schemes–see Christophe Andrieu, Arnaud Doucet, Yıldırım, et al. (2020) for an attempt at implementing this point of view.

We limit probabilistic arguments and notation to a minimum and, in contrast with accepted common wisdom, most often use lower case fonts for both random variables and their realizations in order to alleviate notation. We hope this does not cause confusion.

1.2 MCMC and involutions in the literature

This work is strongly influenced by Tierney (1998) where the possibility of using involutions as "deterministic proposals" is suggested, but not developed as a unifying tool as in the present paper, and the treatment of densities therein is the direct source of inspiration for our own treatment. The papers Fang, J.-M. Sanz-Serna, and Skeel (2014) and Campos and J. Sanz-Serna (2015) were complementary, and revealed to us the importance and generality of the involution point of view, both in the reversible and nonreversible setups, although not always in an explicit manner. A statement of the main abstract result (Theorem 3) was given in Christophe Andrieu and Livingstone (2019, Proposition 3.5) and presented in a series of lectures organized at the Higher School of Economics lectures in St. Petersburg in August 2019 (Christophe Andrieu, 2019), together with various applications, while a preliminary version of the results concerned with NUTS were presented at BayesComp 2020 in Florida in January 2020. We have recently become aware of Graham (2018, p. 64) where the possibility of using an involution as an update was suggested, drawing on an analogy to Green (1995), but not developed. In fact the involutive framework underpins Green (1995) but is not made explicit. The term "Involutive MCMC", perhaps a tautology, was coined in Neklyudov et al. (2020) where classical algorithms are revisited in turn following this perspective, but no connection to earlier literature was made; we also note Cusumano-Towner, Lew, and Mansinghka (2020) with earlier claims and the interesting very recent contribution by Glatt-Holtz, Krometis, and Mondaini (2020). Thin, Kotelevskii, et al. (2020) exploit this type of representation of the MH kernel to design normalising flows and Thin, Durmus, et al. (2020) establish necessary conditions mirroring Tierney (1998) in the skew detailed balance scenario, but also general conditions ensuring aperiodicity and periodicity.

1.3 Notation and definitions

- All real-valued functions we consider are Borel measurable.

- If μ is a measure on $(E, \mathscr{E})$ and $f : \mathsf{E} \to \mathbb{R}$ is a μ-integrable function then we denote the integral $\mu(f) := \int_E f(x)\mu(\mathrm{d}x)$.
- $\min\{a, b\} = a \wedge b$, $\max\{a, b\} = a \vee b$.
- $f \cdot g$ is pointwise product $f \cdot g = x \mapsto f(x)g(x)$, $f/g = x \mapsto f(x)/g(x)$.
- For a set $A \subset E$, the function $\mathbf{1}_A$ is the indicator function of set A, i.e.

$$\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise.} \end{cases}$$

  We also use the notation $\mathbb{I}\{x \in A\} := \mathbf{1}_A(x)$ when the definition of A is explicit and long.
- $\mathbf{1}$ used to denote the constant function $x \mapsto \mathbf{1}$, usage is clear from context.
- For a given x, δ_x is the Dirac measure at x: $\delta_x(A) = \mathbf{1}_A(x)$.
- If $(E, \mathscr{E})$ and $(F, \mathscr{F})$ are measurable spaces, the product measurable space is $(E \times F, \mathscr{E} \otimes \mathscr{F})$ where $\mathscr{E} \otimes \mathscr{F}$ is the product σ-algebra $\sigma(\{A \times B : A \in \mathscr{E}, B \in \mathscr{F}\})$. If μ is a measure on $(E, \mathscr{E})$ and ν a measure on $(F, \mathscr{F})$ then their product measure on $(E \times F, \mathscr{E} \otimes \mathscr{F})$ is $\mu \otimes \nu$ where $(\mu \otimes \nu)(A, B) = \mu(A)\nu(B)$ and define recursively $\mu^{\otimes n} = \mu^{\otimes(n-1)} \otimes \mu$ for $n \in \mathbb{N}_*$.
- If μ is a measure on $(E, \mathscr{E})$ then the restriction of μ to $C \in \mathscr{E}$ is a measure μ_C on $(E, \mathscr{E})$ satisfying $\mu_C(A) := \mu(A \cap C)$ for any $A \in \mathscr{E}$.
- If $\mu(\mathrm{d}x, \mathrm{d}y)$ is a probability measure, we write μ_x to refer to a conditional probability measure for Y given $X = x$. (Polish space)
- A cycle of two Markov kernels $P : E \times \mathscr{E} \to [0, 1]$ and $Q : E \times \mathscr{E} \to [0, 1]$ is the Markov kernel

$$PQ(x, A) = \int P(x, \mathrm{d}y)Q(y, A), \qquad x \in E, A \in \mathscr{E}.$$

- We adopt the standard conventions for products and sums that for $b < a$, $\prod_{i=a}^{b} \cdot = 1$ and $\sum_{i=a}^{b} \cdot = 0$ whatever the nature of the argument.
- For $x \in \mathbb{R}$, $\mathrm{sgn}(x) \in \{-1, 0, 1\}$ is the sign of x.
- We define $\mathbb{N} = \{0, 1, \ldots, \}$ and $\mathbb{N}_* = \{1, 2, \ldots\}$.
- We define $[\![i, j]\!] = \{i, i + 1, \ldots, j\}$ for integers $i \leq j$, and $[\![i]\!] = [\![1, i]\!]$ for $i \in \mathbb{N}_*$.

2 Motivating example

Assume one is interested in sampling from a probability distribution π, defined on some probability space $(\mathsf{Z}, \mathscr{Z})$. A Markov chain Monte Carlo (MCMC) algorithm consists of simulating a realization $\{Z_i; i \geq 0\}$ of a Markov chain such that

$$\mathbb{P}(Z_n \in A) \to \pi(A), \qquad A \in \mathscr{Z},$$

as $n \to \infty$ and/or for functions $f \in L_1(\mathsf{Z}, \pi)$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(Z_i) = \pi(f),$$

One of the fundamental properties required to ensure the above is that, with P denoting the transition probability of the Markov chain, π is left invariant by P. That is, the "global balance" condition holds:

$$\int \pi(\mathrm{d}z)P(z,A) = \pi(A), \qquad z \in \mathsf{Z}, A \in \mathscr{Z}. \tag{1}$$

It is very difficult to verify (1) directly, complicating the design of Markov kernels satisfying this property. A successful approach often consists instead of verifying the stronger, local property of "detailed balance" or $\pi-$reversibility.

Definition 1 (Reversible Markov kernel)**.** For a finite measure μ on $(E, \mathscr{E})$, a Markov kernel $P : E \times \mathscr{E} \to [0,1]$ is μ-reversible if the measures $\mu(\mathrm{d}\xi)P(\xi, \mathrm{d}\xi')$ and $\mu(\mathrm{d}\xi')P(\xi', \mathrm{d}\xi)$ are identical. That is, if,

$$\int_A \mu(\mathrm{d}\xi)P(\xi, B) = \int_B \mu(\mathrm{d}\xi)P(\xi, A), \qquad A, B \in \mathscr{E}.$$

It is straightforward to deduce that (1) holds if P is π-reversible by taking $A = E$ in the definition.

Remark 1. The definition of μ-reversibility is equivalent to: for all measurable $F, G : E \to [0,1]$,

$$\int F(\xi)G(\xi')\mu(\mathrm{d}\xi)P(\xi, \mathrm{d}\xi') = \int G(\xi)F(\xi')\mu(\mathrm{d}\xi)P(\xi, \mathrm{d}\xi'). \tag{2}$$

In particular, we recover the definition with $F = \mathbf{1}_A$ and $G = \mathbf{1}_B$, and for the other direction, we use the identity $\mu(\mathrm{d}\xi)P(\xi, \mathrm{d}\xi') = \mu(\mathrm{d}\xi')P(\xi', \mathrm{d}\xi)$.

Metropolis–Hastings (MH) kernels are a flexible class of $\pi-$reversible Markov kernels for which simulation of the corresponding Markov chain can often be implemented on a computer. A textbook derivation is as follows. Assume that $\mathsf{Z} = \mathbb{R}^d$ and let $\{Q(z,\cdot), z \in \mathsf{Z}\}$ be a family of probability distributions on $(\mathsf{Z}, \mathscr{Z})$ from which it is easy to sample. Assume for presentational simplicity that for any $z \in \mathsf{Z}$, π and $Q(z,\cdot)$ have strictly positive densities with respect to the Lebesgue measure, denoted ϖ and $q(z,\cdot)$. The MH kernel defined by π and Q is given by

$$P(z, \mathrm{d}z') = \alpha(z,z')q(z,z')\mathrm{d}z' + s(z)\delta_z(\mathrm{d}z'),$$

where $\alpha(z,z') = 1 \wedge r(z,z')$, $s(z) = 1 - \int \alpha(z,z')q(z,z')\mathrm{d}z'$ and

$$r(z,z') = \frac{\varpi(z')q(z',z)}{\varpi(z)q(z,z')}.$$

Letting $\rho(z,z') := \varpi(z)q(z,z')$, verifying $\pi-$reversibility can be reduced to checking that for $f, g \colon \mathsf{Z} \to [0,1]$

$$\int f(z)g(z')\rho(z,z')\alpha(z,z')\mathrm{d}z\mathrm{d}z' = \int g(z)f(z')\rho(z,z')\alpha(z,z')\mathrm{d}z\mathrm{d}z', \tag{3}$$

since

$$\int f(z)g(z')\varpi(z)s(z)\delta_z(\mathrm{d}z')\mathrm{d}z = \int f(z)g(z)\varpi(z)s(z)\mathrm{d}z = \int g(z)f(z')\varpi(z)s(z)\delta_z(\mathrm{d}z')\mathrm{d}z.$$

It is a standard exercise to show that $\rho(z,z')\alpha(z,z') = \rho(z',z)\alpha(z',z)$ and conclude that (3) holds. We outline now a less direct way, which however has the benefit of highlighting important generic properties required.

Define $\xi := (z,z')$, $\mathrm{d}\xi = \mathrm{d}z\mathrm{d}z'$, $\phi(z,z') = (z',z)$ and $F_0(z,z') = f(z)$ and $G_0(z,z') = g(z)$ then (3) can be re-expressed as

$$\int F_0(\xi)G_0 \circ \phi(\xi)\rho(\xi)\alpha(\xi)\mathrm{d}\xi = \int F_0 \circ \phi(\xi)G_0(\xi)\rho(\xi)\alpha(\xi)\mathrm{d}\xi. \tag{4}$$

Further notice that the acceptance ratio is of the form $r(\xi) = \rho\circ\phi/\rho(\xi)$ and that, using that $\phi\circ\phi = \mathrm{Id}$,

$$r \circ \phi(\xi) = \frac{\rho \circ \phi}{\rho} \circ \phi(\xi) = \frac{\rho}{\rho \circ \phi}(\xi) = \frac{1}{r}(\xi),$$

therefore implying

$$r(\xi)\alpha \circ \phi(\xi) = r(\xi)\left[1 \wedge r \circ \phi(\xi)\right] = \alpha(\xi). \tag{5}$$

We now show that (4) holds for any measurable $F, G : \mathsf{Z}^2 \to [0, 1]$

$$\begin{aligned}\int F(\xi)G \circ \phi(\xi)\rho(\xi)\alpha(\xi)\mathrm{d}\xi &= \int F(\xi)G \circ \phi(\xi)\rho(\xi)r(\xi)\alpha \circ \phi(\xi)\mathrm{d}\xi \\ &= \int F(\xi)G \circ \phi(\xi)\rho \circ \phi(\xi)\alpha \circ \phi(\xi)\mathrm{d}\xi \\ &= \int F \circ \phi(\xi')G(\xi')\rho(\xi')\alpha(\xi')\mathrm{d}\xi',\end{aligned}$$

where we have used $\alpha(\xi) = r(\xi)\alpha \circ \phi(\xi)$, $r(\xi) = \rho \circ \phi/\rho(\xi)$ the change of variable $\xi' = \phi(\xi)$ and the fact that ϕ is an involution with Jacobian $|\mathrm{det}\phi'(\xi)| = 1$ (see Theorem 4). This therefore implies (3) and in turn that P is $\pi-$reversible. In fact, letting $\mu(\mathrm{d}\xi) := \rho(\xi)\mathrm{d}\xi$, we notice that this establishes $\mu-$reversibility of an MH kernel targeting the extended probability distribution μ.

This presentation has the advantage of highlighting a set of generic properties sufficient to establish $\pi-$reversibility:

(a) the distribution π is a marginal of a probability distribution μ,

(b) the proposed state is obtained by applying an involution ϕ to ξ,

(c) it holds that $\alpha(\xi)\mu(\mathrm{d}\xi) = \alpha \circ \phi(\xi)\mu^{\phi}(\mathrm{d}\xi)$ with μ^{ϕ} the probability distribution of $\xi' = \phi^{-1}(\xi) = \phi(\xi)$,

suggesting that more general choices of μ, ϕ and α can also define $\pi-$reversible Markov kernels. It can be shown (Theorem 4) that the first two properties automatically imply the mathematical existence of α such that the third property holds, highlighting the fundamental rôle played by the involutory nature of ϕ. Practical implementation of the algorithm requires two additional properties of μ: the existence of a tractable probability density to compute α and ease of sampling from the conditional distribution in $\mu(\mathrm{d}\xi) = \pi(\mathrm{d}\xi_0)\mu_{\xi_0}(\mathrm{d}\xi_{-0})$.

The clear benefit of this approach is that establishing correctness becomes a purely mechanical, or "algebraic", exercise, therefore improving clarity of arguments and facilitating communication.

3 General scenario

In order to gain generality and clarify we will appeal to a very small number of standard measure theoretical notions and results related to change of variables and Radon–Nykodim derivatives. Although it is always a good idea to check the proof of classical results, there is no need to do so in order to understand the content of this manuscript.

Definition 2 (Pushforward)**.** Let μ be a measure on $(E, \mathscr{E})$ and $\varphi : (E, \mathscr{E}) \to (F, \mathscr{F})$ a measurable function. The pushforward of μ by φ is defined by

$$\mu^{\varphi}(A) = \mu(\varphi^{-1}(A)), \qquad A \in \mathscr{F},$$

where $\varphi^{-1}(A) = \{x \in E : \varphi(x) \in A\}$ is the preimage of A under φ.

For example, if μ is a probability distribution then μ^φ is the probability measure associated with $\varphi(X)$ when $X \sim \mu$.

Definition 3 (Dominating and equivalent measures)**.** For two measures μ and ν on the same measurable space $(E, \mathscr{E})$,

(a) μ is said to dominate ν if for all measurable $A \in \mathscr{E}$, $\nu(A) > 0 \Rightarrow \mu(A) > 0$ – this is denoted $\mu \gg \nu$.

(b) μ and ν are equivalent, written $\mu \equiv \nu$, if $\mu \gg \nu$ and $\nu \gg \mu$.

We will need the notion of Radon-Nikodym derivative:

Theorem 1 (Radon–Nikodym)**.** *Let μ and ν be σ-finite measures on $(E, \mathscr{E})$. Then $\nu \ll \mu$ if and only if there exists an essentially unique, measurable, non-negative function f such that*

$$\int_A f(\xi)\mu(\mathrm{d}\xi) = \nu(A), \qquad A \in \mathscr{E}.$$

Therefore we can view $\mathrm{d}\nu/\mathrm{d}\mu := f$ as the density of ν w.r.t μ and in particular if g is integrable w.r.t. ν then

$$\int g(\xi)\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(\xi)\mu(\mathrm{d}\xi) = \int g(\xi)\nu(\mathrm{d}\xi).$$

This is covered by Billingsley (1995, Theorems 32.2 & 16.11).

If μ is a measure and f a non-negative, measurable function then $\mu \cdot f$ is the measure $(\mu \cdot f)(A) = \int \mathbf{1}_A(x) f(x)\mu(\mathrm{d}x)$, i.e. the measure $\nu = \mu \cdot f$ such that the Radon–Nikodym derivative of $\mathrm{d}\nu/\mathrm{d}\mu = f$.

Theorem 2 (Change of variables)**.** *A function $f : F \to \mathbb{R}$ is integrable w.r.t. μ^φ if and only if $f \circ \varphi$ is integrable w.r.t. μ, in which case*

$$\int_F f(\xi)\mu^\varphi(\mathrm{d}\xi) = \int_E f \circ \varphi(\xi)\mu(\mathrm{d}\xi). \tag{6}$$

This can be found in Billingsley (1995, Theorem 16.13).

3.1 An abstract result

The following result is central to the design of MH-based MCMC, formalizes the observations made in Section 2 and generalizes parts of Tierney (1998, Proposition 1 and Theorem 2), concerned with the specific involution $\phi(z, z') = (z', z)$ and a particular form of distribution μ. We do not pursue necessity conditions here, to keep the presentation brief and focused on practical consequences: Tierney (1998) discusses such issues, while Thin, Durmus, et al. (2020) revisits these issues in a particle nonreversible setup (see Section 4). The proof can be found in Appendix A. This result mirrors Christophe Andrieu and Livingstone (2019, Proposition 2).

Theorem 3. *Let μ be a finite measure on $(E, \mathscr{E})$, $\phi : E \to E$ an involution. Then*

(a) there exists a set $S = S(\mu, \mu^\phi) \in \mathscr{E}$ such that

(i) $\phi(S) = S$,

(ii) with $\mu_S(A) := \mu(A \cap S)$ for any $A \in \mathscr{E}$ we have $\mu_S^\phi \equiv \mu_S$,

(iii) μ *and* μ^ϕ *are mutually singular on* $S^{\complement}$*, i.e. there exist sets* $A, B \in \mathscr{E}$ *such that* $A \cap B = \emptyset$*,* $A \cup B = S^{\complement}$ *and* $\mu(A) = \mu^\phi(B) = 0$*.*

(b) defining for $\xi \in \Xi$*,*

$$r(\xi) := \begin{cases} \mathrm{d}\mu_S^\phi / \mathrm{d}\mu_S(\xi) & \xi \in S, \\ 0 & \text{otherwise}, \end{cases} \tag{7}$$

and letting $a : [0, \infty) \to [0, 1]$ *such that*

$$a(r) = \begin{cases} 0 & r = 0 \\ ra(1/r) & r > 0 \end{cases},$$

we have that,

(i) for $\xi \in \Xi$*,*

$$\alpha(\xi) := a \circ r(\xi) = \begin{cases} r(\xi) \cdot \alpha \circ \phi(\xi) & \xi \in S, \\ 0 & \text{otherwise}, \end{cases}$$

(ii) for any measurable $F, G : E \to [0, 1]$*,*

$$\int F(\xi) G \circ \phi(\xi) \alpha(\xi) \mu(\mathrm{d}\xi) = \int F \circ \phi(\xi) G(\xi) \alpha(\xi) \mu(\mathrm{d}\xi),$$

(iii) the Markov kernel Π *defined by*

$$\Pi(\xi, \{\phi(\xi)\}) = \alpha(\xi) = 1 - \Pi(\xi, \{\xi\}),$$

is μ*-reversible.*

Remark 2. The condition on a is satisfied by $a(r) = 1 \wedge r$ (corresponding to the Metropolis–Hastings acceptance probability), and also $a(r) = r/(1 + r)$ (Barker's acceptance probability; see Example 3), therefore ensuring the existence of Π and P.

In practice one is interested in the component ξ_0 of μ, which is distributed according to π. In fact, the Markov kernel Π in Theorem 3 can be used to define a π-invariant Markov kernel P. The proof can be found in Appendix A.

Proposition 1. *Let* π *be a probability distribution on* $(\mathsf{Z}, \mathscr{Z})$ *and let* μ *be a probability distribution on* $(E, \mathscr{E})$ *such that*

$$\mu(\mathrm{d}\xi) := \pi(\mathrm{d}\xi_0) \mu_{\xi_0}(\mathrm{d}\xi_{-0}),$$

where μ_{ξ_0} *denotes the conditional distribution of* ξ_{-0} *given* ξ_0 *under* μ*. Then the Markov kernel*

$$P(\xi_0, A) := \int \mathbf{1}_A(\xi_0') \mu_{\xi_0}(\mathrm{d}\xi_{-0}) \Pi(\xi; \mathrm{d}\xi'), \qquad A \in \mathscr{Z},$$

is π*-reversible.*

An algorithmic description of P is given in Alg. 1 highlighting the practical requirement that sampling from $\mu_{\xi_0}(\cdot)$ for $\xi_0 \in \mathsf{Z}$ should be tractable.

The implication of these results should be clear. If sampling from π is of interest, any choice of μ of the form

$$\mu(\mathrm{d}\xi) = \pi(\mathrm{d}\xi_0) \mu_{\xi_0}(\mathrm{d}\xi_{-0}), \tag{8}$$

together with an involution ϕ and an acceptance function a defines a $\pi-$reversible Markov kernel/chain. It turns out that all MH-type kernels we are aware of, including advanced and complex implementations, can be described and immediately justified using this framework.

Algorithm 1 To sample from $P(\xi_0, \cdot)$

(a) Given ξ_0 , sample $\xi_{-0} \sim \mu_{\xi_0}$,

(b) Compute $\xi' = \phi(\xi)$,

(c) With probability $\alpha(\xi)$ return ξ'_0, otherwise return ξ_0.

Remark 3. The framework specified is very flexible: to define a π-reversible Markov kernel P, whose simulation is described in Algorithm 1, it is sufficient to define a triple (μ, ϕ, a) such that π is the ξ_0-marginal of μ. This is analogous to the definition of a traditional Metropolis–Hastings kernel via the choice (π, Q) in Section 2. Importantly the nature of ξ_{-0} is *a priori* arbitrary and does not have to coincide with that of ξ_0, therefore providing great freedom. In general, the association is not unique: there are several (μ, ϕ, a) triples corresponding to the same Markov kernel P. In the sequel we will focus primarily on the the measure-involution pair (μ, ϕ), since the choice of a can be taken independently of the choice of (μ, ϕ) from a theoretical perspective.

In the sequel we will consider Markov kernels Π as in Theorem 3, or derivatives such as P in Proposition 1 as Metropolis–Hastings type kernels.

Remark 4. In the context of Proposition 1 it is natural to ask whether theoretical properties, such as optimality in terms of optimal variance of Π translate into optimality for P. The answer is yes and follows by application of the results of Maire, Douc, and Olsson (2014), later extended in Christophe Andrieu and Livingstone (2019) to the nonreversible scenario treated in Section 4.

We now provide examples of commonly used Markov kernels, which can be recognized by the particular form of μ and ϕ, and for which expressions of the corresponding acceptance ratios is left to Section 3.2. This highlights the fact that the acceptance ratio is a function depending only on μ and ϕ.

Example 1. The textbook presentation of the MH kernel considered in the introduction corresponds to the choice of a family of conditional probability distributions $\{\mu_z(\cdot) = Q(z, \cdot), z \in \mathsf{Z}\}$ on $(\mathsf{Z}, \mathscr{Z})$, $\xi = (z, z') \in E = \mathsf{Z} \times \mathsf{Z}$, $\phi(z, z') = (z', z)$ and $a = r \mapsto 1 \wedge r$.

Example 2. The Random Walk Metropolis (RWM) can be thought of as corresponding to the choice $\{\mu_z(\cdot) = \kappa(\cdot), z \in \mathsf{Z}\}$ for some probability distribution κ on $(\mathsf{Z}, \mathscr{Z})$, $\xi = (z, v)$ and $\phi(\xi) = (z + v, -v)$. Alternatively, one may express the RWM as a special case of Example 1 so that $\{\mu_z(\cdot) = Q(\cdot), z \in \mathsf{Z}\}$, $\xi = (z, z')$ and $\phi(z, z') = (z', z)$.

Example 3 (Metropolis–Hastings, Barker, etc.)**.** Let μ be as in Example 1, and let $\xi = (z, z')$ with $\xi_0 = z$. Then Alg 1 corresponds to simulating from the Metropolis–Hastings (resp. Barker) kernel when $\alpha = a \circ r$, with $a(v) = 1 \wedge v$ (resp. $a(v) = 1/(1 + v)$). This corresponds to the presentation adopted by Tierney (1998) and commonly adapted in the literature.

The requirement that ϕ be an involution may appear restrictive, but in fact for a given invertible function one can define a corresponding involution by extending the space.

Remark 5. Let μ be a measure admitting π as a marginal and $\phi : E \to E$ be invertible, but not an involution. Then (μ_0, ϕ_0) is a corresponding measure-involution pair, where $\mu_0(\mathrm{d}\xi, \mathrm{d}v) := \mu(\mathrm{d}\xi)\mathbb{I}\{v \in \{-1, 1\}\}/2$ and $\phi_0(\xi, v) := (\phi^v(\xi), -v)$ on $E_0 := E \times \{-1, 1\}$. Since μ_0 admits μ as a marginal, it also admits π as a marginal.

Example 4 (Ordered overrelaxation (Neal, 1998))**.** A Gibbs sampler can be thought of as a MH update where conditional distributions of the target distribution π on $(\mathsf{Z}, \mathscr{Z})$ are used in the proposal mechanism. To fix ideas assume $\mathsf{Z} = \mathsf{X} \times \mathsf{Y}$ where $\mathsf{Y} \subset \mathbb{R}$ and let η be one such conditional distribution

on $(\mathsf{Y}, \mathscr{Y})$ from which sampling is tractable. The goal of the method is to develop a numerical implementation the following remark. Let $(x, y) \in \mathsf{Z}$ and let F_η be the cumulative distribution function (cdf) corresponding to η, where x is implicit. Then $y' = F_\eta^{-1}\big(1 - F_\eta(y)\big)$ is antithetic to y–in fact for $u \sim \mathrm{Uniform}(0,1)$ the pair $\big(F_\eta^{-1}(u), F_\eta^{-1}\big(1-u\big)\big)$ is the lower bound in the Fréchet class of bivariate distributions of marginals η. The numerical approximation of this remark exploits the link between empirical cdf and order statistics. One can sample multiple times independently from η, leading to the probability distribution, for $n \in \mathbb{N}_*$, on $(\mathsf{Z} \times \mathsf{Y}^n, \mathsf{Z} \otimes \mathsf{Y}^{\otimes n})$

$$\mu = \pi \otimes \eta^{\otimes n}.$$

Let $z = (x, y_0)$ and let $\sigma \colon [\![0, n]\!] \to [\![0, n]\!]$ be the $\xi := (z, y_1, \ldots, y_n)$-dependent permutation such that

$$y_{\sigma(0)} \leq \cdots \leq y_{\sigma(n)}$$

and let $r \in [\![0, n]\!]$ be the integer such that $\sigma(r) = 0$ i.e. y_0 is the $r-$th rank order statistic among $y_0, y_1, \ldots, y_n$. Now we consider the following involution, for $r \in [\![2, n-1]\!]$

$$\phi(z, y_1 \ldots, y_n) = (x, y_{\sigma(n-r)}, y_1, \ldots, y_{\sigma(n-r)-1}, y_0, y_{\sigma(n-r)+1}, \ldots, y_n),$$

with straightforward adaptation if $r \in \{0, 1, n\}$. It should be clear, from the exchangeability conditional upon $x \in \mathsf{X}$, that for $\xi \in S\big(\mu, \mu^\phi\big)$, $r(\xi) = 1$. One can naturally replace η with a proposal distribution of our choosing, but the acceptance ratio is then not identically equal to 1.

Adopting this point of view makes establishing reversibility routine, even in complex scenarios. However practical implementation of the update requires an explicit expression for the acceptance ratio r in (7), not provided by the results above.

Remark 6. Alg. 1 is conceptually simple, but in practice it may be expedient to avoid a direct implementation. What is actually required to simulate from $P(\xi_0, \cdot)$ is to sample a Bernoulli$(\alpha(\xi))$ random variable, where $\xi_{-0} \sim \mu_{\xi_0}$ and to compute $\phi(\xi)_0$. In particular, it may not be necessary to simulate or store ξ_{-0} in its entirety to perform these task, e.g. when ξ_{-0} is large or even infinite-dimensional. Some examples are provided in Section 4.

We will primarily focus on Alg. 1 in the sequel. Hence, for examples and applications of this framework we will identify an appropriate (μ, ϕ), hence defining Π in Theorem 3 up to the choice of a. The corresponding π-reversible Markov kernel is then defined by P in Proposition 1. There are, of course, other μ-invariant kernels that can be constructed using Π. For example, letting R define the refreshment kernel

$$R(\xi, \mathrm{d}\xi') = \delta_{\xi_0}(\mathrm{d}\xi_0')\mu_{\xi_0}(\mathrm{d}\xi_{-0}'),$$

Alg. 1 corresponds to tracking the ξ_0-coordinate of $R\Pi(\xi_0, \cdot)$. One could instead define a μ-invariant kernel as $\gamma R + (1-\gamma)\Pi$ for some $\gamma \in (0,1)$. Even more generally, one could replace R with another Markov kernel that only leaves the conditional distribution μ_{ξ_0} invariant. The cycle $R\Pi$ is then μ-invariant and would sometimes be referred to as a Metropolis-within-Gibbs (MwG) kernel, although we note that in this case the corresponding ξ_0-coordinate of the μ-invariant Markov chain would in general not be Markov. More generally we will refer to an algorithm involving a mixture ("random-scan") or cycle ("deterministic scan") of kernels targetting the same distribution as a MwG, a widely accepted misnomer.

3.2 Densities and the acceptance ratio

In order to compute the acceptance ratio r in Theorem 3, one must identify S and have an expression for $\mathrm{d}\mu_S^\phi / \mathrm{d}\mu_S$. We show below how to phrase these objects in terms of a density $\rho = \mathrm{d}\mu/\mathrm{d}\lambda$, where λ is an appropriate reference measure. Such a density is often available *a priori* in practice.

Proposition 2. *Let μ be a finite measure on $(E, \mathscr{E})$, $\phi : E \to E$ an involution, let $\lambda \gg \mu$ be a σ-finite measure satisfying $\lambda \equiv \lambda^\phi$ and let $\rho = \mathrm{d}\mu/\mathrm{d}\lambda$. Then we can take $S = S(\mu, \mu^\phi)$ to be $S = \{\xi : \rho(\xi) \wedge \rho \circ \phi(\xi) > 0\}$ and*

$$r(\xi) = \begin{cases} \frac{\rho \circ \phi}{\rho}(\xi) \frac{\mathrm{d}\lambda^\phi}{\mathrm{d}\lambda}(\xi) & \xi \in S, \\ 0 & \text{otherwise}, \end{cases} \tag{9}$$

in Theorem 3.

The proof can be found in Appendix B.2. In many situations λ will be the Lebesgue or counting measure, but can also be a product of these, or an infinite-dimensional probability measure such as a Gaussian measure Hairer, Stuart, Vollmer, et al. (2014) or the law of a Markov chain (this is treated in Subsection 5.2). Computing (9) involves additionally computing the density $\mathrm{d}\lambda^\phi/\mathrm{d}\lambda$.

Remark 7. If in Proposition 2 λ is invariant under ϕ, i.e. $\lambda = \lambda^\phi$ then $r = \rho \circ \phi / \rho$. In theory, it is always possible to find a reference measure invariant under ϕ, e.g. one could instead of λ take $\lambda_0 := \lambda + \lambda^\phi$ or even $\lambda_0 := \mu + \mu^\phi$, which underpins the proof of Theorem 3. However, it may not be straightforward or natural to compute the density $\mathrm{d}\mu/\mathrm{d}\lambda_0$, while there is often a natural choice of λ for which $\mathrm{d}\mu/\mathrm{d}\lambda$ can be computed.

A standard scenario is when λ is the Lebesgue measure on $E = \mathbb{R}^d$ and ϕ is a diffeomorphism, in which case $\mathrm{d}\lambda^\phi/\mathrm{d}\lambda$ corresponds to the absolute value of the determinant of the Jacobian, since then for any λ-integrable f (see Theorem B.2 in Appendix B)

$$\int f \circ \phi(\xi) \left|\mathrm{det}\phi'(\xi)\right| \lambda(\mathrm{d}\xi) = \int f(\xi) \lambda(\mathrm{d}\xi) = \int f \circ \phi(\xi) \lambda^\phi(\mathrm{d}\xi),$$

while for an arbitrary, measurable, non-negative $g : E \to \mathbb{R}$ we can take $f = g \circ \phi^{-1}$ to obtain $g = f \circ \phi$ and hence,

$$\int g(\xi) \left|\mathrm{det}\phi'(\xi)\right| \lambda(\mathrm{d}\xi) = \int g(\xi) \lambda^\phi(\mathrm{d}\xi).$$

The example of the introduction corresponds to this scenario, but where in addition ϕ is an involution and the reference measure is invariant under ϕ.

Remark 8. There are several ways one can determine $\mathrm{d}\lambda^\phi/\mathrm{d}\lambda$ in common situations. For example:

(a) Let $E = \mathbb{R}^d$ with $\xi = (z_1, \ldots, z_d)$, and ϕ be an involution that permutes its input, i.e. $\phi(z_1, \ldots, z_d) = (z_{\sigma(1)}, \ldots, z_{\sigma(d)})$ for some permutation σ of $\{1, \ldots, d\}$. Then since $\phi'(\xi)$ is the corresponding permutation matrix and all permutations have a determinant in $\{-1, 1\}$, we obtain $|\mathrm{det}\phi'(\xi)| = 1$. So if λ is the Lebesgue measure on $\mathbb{R}^d$ then $\lambda^\phi = \lambda$.

(b) Let μ be a measure with countable support X, and let λ be the counting measure on $E = \mathsf{X} \cup \phi(\mathsf{X}) = \mathsf{X} \cup \{\phi(x) : x \in \mathsf{X}\}$. Then for an arbitrary, measurable $A \subseteq E$ we have $\lambda^\phi(A) = \lambda(\phi^{-1}(A)) = |A| = \lambda(A)$ since ϕ is an involution. Hence $\lambda = \lambda^\phi$ so $\mathrm{d}\lambda^\phi/\mathrm{d}\lambda = 1$.

In some of our applications, μ has continuous and discrete components, and a density with respect to a product of a Lebesgue measure and a counting measure. When the involution for the discrete component does not depend on the continuous component, we have the following result.

Lemma 1. *Let λ_X be the Lebesgue measure on X, λ_Y the counting measure on Y and $g : \mathsf{Y} \to \mathsf{Y}$ be an involution with $g(\mathsf{Y}) \subseteq \mathsf{Y}$. Let $f : \mathsf{X} \times \mathsf{Y} \to \mathsf{X}$ be a function such that*

$$\phi(x, y) = (f(x, y), g(y)),$$

is an involution. Then $\mathrm{d}\lambda^\phi/\mathrm{d}\lambda = \left|\mathrm{det} f'_y(x)\right|$.

We are now in a position to provide expressions for the acceptance ratios in Examples 1–2.

Example 5 (Metropolis–Hastings acceptance ratio)**.** Let π and $\{Q(z,\cdot), z \in \mathsf{Z}\}$ be probability measures on $(\mathsf{Z}, \mathscr{Z})$ such that with ν the Lebesgue or counting measure we have $\nu \gg \pi$ and $\nu \gg Q(z,\cdot)$ for each $z \in \mathsf{Z}$. Let $\varpi(z) = \mathrm{d}\pi/\mathrm{d}\nu(z)$ and $q(z,z') = \mathrm{d}Q(z,\cdot)/\mathrm{d}\nu(z')$ for all $(z,z') \in \mathsf{Z}^2$. With $\xi = (z,z')$ we let $\mu(\mathrm{d}\xi) = \pi(\mathrm{d}z)Q(z,\mathrm{d}z')$, and $\phi(z,z') = (z',z)$. Then with $\lambda^\phi = \lambda := \nu \times \nu$ we obtain $\rho(\xi) = \varpi(z)q(z,z')$ and $\rho \circ \phi(\xi) = \varpi(z')q(z',z)$ and the acceptance ratio is, for $\xi \in S(\mu,\mu^\phi) = \{\xi : \rho(\xi) \wedge \rho \circ \phi(\xi) > 0\}$,

$$r(\xi) = \frac{\rho \circ \phi(\xi)}{\rho(\xi)} = \frac{\varpi(z')q(z',z)}{\varpi(z)q(z,z')}.$$

Example 6 (Random walk Metropolis ratio)**.** The setup is similar to above but we assume that $\mathsf{Z} = \mathbb{R}^d$, ν is the Lebesgue measure, $q(z,v) = q(v) := \mathrm{d}Q/\mathrm{d}\nu(v)$ for $(z,v) \in \mathsf{Z}^2$ and $q(v) = q(-v)$ for $v \in \mathsf{Z}$. Here $\lambda = \nu \times \nu$, $\xi = (z,v) \in \mathsf{Z}^2$, $\phi(\xi) = (z+v,-v)$ and $|\mathrm{det}\phi'(\xi)| = 1$, leading to

$$r(\xi) = \frac{\rho \circ \phi(\xi)}{\rho(\xi)} = \frac{\varpi(z+v)q(-v)}{\varpi(z)q(v)} = \frac{\varpi(z+v)}{\varpi(z)}.$$

It is possible to consider the setting where $\xi_0 = \xi$ and ϕ is an involution with non-unit Jacobian. Such situations are related, e.g., to the Monte Carlo Markov kernels based on deterministic transformations proposed by Dutta and Bhattacharya (2014).

Example 7. Assume $\rho = \mathrm{d}\mu/\mathrm{d}\lambda$ with $\{\xi \in E : \rho(\xi) > 0\} = (0,1)$ with λ the Lebesgue measure on $\mathbb{R}$ and let $\phi(\xi) = 1/(2\xi)$. One can deduce that λ and λ^ϕ are equivalent with $\mathrm{d}\lambda^\phi/\mathrm{d}\lambda(\xi) = |\phi'(\xi)| = 1/(2\xi^2)$. We obtain $S = S(\mu,\mu^\phi) = \{\xi \in E : \rho(\xi) \wedge \rho \circ \phi(\xi) > 0\} = (0,1) \cap \phi^{-1}(0,1) = [1/2,1)$. Therefore $r(\xi) = \rho \circ \phi(\xi)/(\rho(\xi)2\xi^2)$ for $\xi \in S$ and $r(\xi) = 0$ otherwise.

Example 8. Consider π a probability measure on $\mathbb{R}$ dominated by Lebesgue and $\varphi(x) = x^3$, which is invertible but not an involution. Then following Remark 5, we can extend the space to $E = \mathbb{R} \times \{-1,1\}$ and define $\xi = (x,k)$, $\mu(\mathrm{d}\xi) = \pi(\mathrm{d}x)\mathbb{I}(k \in \{-1,1\})/2$, λ to be the product of the Lebesgue measure and the counting measure, and $\phi(\xi) = \phi(x,k) = (\varphi^k(x),-k)$. Following Lemma 1, we obtain

$$\frac{\mathrm{d}\lambda^\phi}{\mathrm{d}\lambda}(x,k) = \left|\mathrm{det}(\varphi^k)'(x)\right| = \begin{cases} 3x^2 & k = 1, \\ \frac{1}{3}|x|^{-2/3} & k = -1. \end{cases}$$

A slightly more general version of Example 5 above can be used when Q is reversible w.r.t. some measure.

Example 9. Let π be probability measures on $(\mathsf{Z}, \mathscr{Z})$, ν be a reference measure such that $\pi \ll \nu$ and assume that Q is ν-reversible. Then with $\xi = (z,z')$ and $\varpi = \mathrm{d}\pi/\mathrm{d}\nu$,

$$\mu(\mathrm{d}\xi) = \pi(\mathrm{d}z)Q(z,\mathrm{d}z') = \varpi(z)\nu(\mathrm{d}z)Q(z,\mathrm{d}z'),$$

that is $\rho(\xi) = \varpi(\xi_0)$ with $\lambda(\mathrm{d}\xi) = \nu(\mathrm{d}\xi_0)Q(\xi_0,\mathrm{d}\xi_{-0})$ and by assumption $\lambda^\phi = \lambda$ for $\phi(z,z') = (z',z)$ for $(z,z') \in \mathsf{Z}^2$. Therefore

$$r(\xi) = \frac{\rho \circ \phi}{\rho}(\xi) = \frac{\varpi(z')}{\varpi(z)}, \qquad \xi \in S = \{(z,z') \in \mathsf{Z}^2 : \varpi(z) \wedge \varpi(z') > 0\}.$$

In many common RWM kernels, $\nu$is the Lebesgue (resp. counting) measure on a continuous (resp. discrete) state space.

Example 10. (Simplified Neal tempering) Let π be a multimodal distribution on $(\mathsf{Z}, \mathscr{Z})$. A strategy proposed by Neal (1996) to mitigate the effect of multimodality on consists of using an instrumental

distribution $\tilde{\pi} \equiv \pi$ also defined on $(\mathsf{Z}, \mathscr{Z})$, related to π but less multimodal, to improve the rate of moves between modes of π. More specifically define

$$\mu(\mathrm{d}\xi) = \pi(\mathrm{d}\xi_0)Q(\xi_0, \mathrm{d}\xi_1)\tilde{Q}(\xi_1, \mathrm{d}\xi_2)Q(\xi_2, \mathrm{d}\xi_3),$$

where $\tilde{Q}$ (resp. Q) is $\tilde{\pi}-$reversible (resp. $\pi-$reversible) and consider the involution on $E = \mathsf{Z}^4$ such that $\phi(\xi_0, \xi_1, \xi_2, \xi_3) = (\xi_3, \xi_2, \xi_1, \xi_0)$. Using these properties, we obtain

$$\begin{aligned}
\mu^{\phi}(\mathrm{d}\xi) &= \pi(\mathrm{d}\xi_3)Q(\xi_3, \mathrm{d}\xi_2)\tilde{Q}(\xi_2, \mathrm{d}\xi_1)Q(\xi_1, \mathrm{d}\xi_0) \\
&= \pi(\mathrm{d}\xi_2)Q(\xi_2, \mathrm{d}\xi_3)\tilde{Q}(\xi_2, \mathrm{d}\xi_1)Q(\xi_1, \mathrm{d}\xi_0) \\
&= \frac{\mathrm{d}\pi}{\mathrm{d}\tilde{\pi}}(\xi_2)Q(\xi_2, \mathrm{d}\xi_3)\tilde{\pi}(\mathrm{d}\xi_2)\tilde{Q}(\xi_2, \mathrm{d}\xi_1)Q(\xi_1, \mathrm{d}\xi_0) \\
&= \frac{\mathrm{d}\pi}{\mathrm{d}\tilde{\pi}}(\xi_2)Q(\xi_2, \mathrm{d}\xi_3)\tilde{Q}(\xi_1, \mathrm{d}\xi_2)\tilde{\pi}(\mathrm{d}\xi_1)Q(\xi_1, \mathrm{d}\xi_0) \\
&= \frac{\mathrm{d}\pi}{\mathrm{d}\tilde{\pi}}(\xi_2)\frac{\mathrm{d}\tilde{\pi}}{\mathrm{d}\pi}(\xi_1)Q(\xi_2, \mathrm{d}\xi_3)\tilde{Q}(\xi_1, \mathrm{d}\xi_2)\pi(\mathrm{d}\xi_1)Q(\xi_1, \mathrm{d}\xi_0) \\
&= \frac{\mathrm{d}\pi}{\mathrm{d}\tilde{\pi}}(\xi_2)\frac{\mathrm{d}\tilde{\pi}}{\mathrm{d}\pi}(\xi_1)Q(\xi_2, \mathrm{d}\xi_3)\tilde{Q}(\xi_1, \mathrm{d}\xi_2)Q(\xi_0, \mathrm{d}\xi_1)\pi(\mathrm{d}\xi_0) \\
&= \frac{\mathrm{d}\pi}{\mathrm{d}\tilde{\pi}}(\xi_2)\frac{\mathrm{d}\tilde{\pi}}{\mathrm{d}\pi}(\xi_1)\mu(\mathrm{d}\xi).
\end{aligned}$$

It follows that we can take $S = \{\xi \in E \colon \mathrm{d}\tilde{\pi}/\mathrm{d}\pi(\xi_1)\, \mathrm{d}\pi/\mathrm{d}\tilde{\pi}(\mathrm{d}\xi_2) > 0\}$, and we obtain,

$$r(\xi) = \frac{\mathrm{d}\tilde{\pi}}{\mathrm{d}\pi}(\xi_1)\frac{\mathrm{d}\pi}{\mathrm{d}\tilde{\pi}}(\xi_2), \qquad \xi \in S.$$

In this case, we can think of $\lambda = \mu \equiv \mu^{\phi} = \lambda^{\phi}$, $\rho \equiv 1$ and $\mathrm{d}\lambda^{\phi}/\mathrm{d}\lambda = r(\xi)$ on S. In practice, computation of the acceptance ratio may be facilitated by convenient densities for $\tilde{\pi}$ and π with respect to a common dominating measure. The above can be viewed as the justification for the tempered transitions kernel introduced by Neal (1996), where several instrumental distributions are used; these ideas are also related to the methodology in Neal (2005).

Example 11 (Penalty method Ceperley and Dewing (1999))**.** In this scenario $\mu\big(\mathrm{d}(z, w)\big) = \mu_0(\mathrm{d}z)Q_z(\mathrm{d}w)$, $\xi = (z, w) \in \mathsf{Z} \times \mathsf{W}$ with $\mathsf{W} \subset \mathbb{R}_+^*$, $\phi(z, w) = \big(\phi_0(z), 1/w\big)$ for $\phi_0 \colon \mathsf{Z} \to \mathsf{Z}$ an involution and $w \cdot Q_z(\mathrm{d}w) = Q^{1/\cdot}_{\phi_0(z)}(\mathrm{d}w)$ for $w > 0$, where for any $f \colon \mathsf{Z} \to [0, 1]$

$$\int f(w)Q_z^{1/\cdot}(\mathrm{d}w) := \int f\big(w^{-1}\big)Q_z(\mathrm{d}w),$$

therefore implying for $\xi \in S$

$$r(\xi) = \frac{\mathrm{d}\pi^{\phi_0}}{\mathrm{d}\pi}(z)\frac{Q^{1/\cdot}_{\phi_0(z)}(\mathrm{d}w)}{Q_z(\mathrm{d}w)} = \frac{\mathrm{d}\pi^{\phi_0}}{\mathrm{d}\pi}(z)w.$$

The motivation for this setup is concerned with the situation where a noisy version of the acceptance ratio $\mathrm{d}\pi^{\phi_0}/\mathrm{d}\pi(z)$ is available, where the noise is additive in the log-domain, corresponding to noisy energies in Physics. The condition on Q_z is satisfied by the random variable $W := \exp(-\sigma_z^2/2 + \sigma_z Z)$

with $Z \sim \mathcal{N}(0,1)$ for $z \mapsto \sigma_z = \sigma_{\phi_0(z)}$ because

$$\begin{aligned}
\int f(w) Q_z(\mathrm{d}w) &= \int f\big[\exp(-\sigma_z^2/2 + \sigma_z x)\big]\mathcal{N}(\mathrm{d}x; 0, 1) \\
&= \int f\big[\exp\big(\sigma_z^2/2 - \sigma_z(x + \sigma_z)\big)\big]\mathcal{N}(\mathrm{d}x; 0, 1) \\
&= \int f\big[\exp(\sigma_z^2/2 - \sigma_z x)\big]\mathcal{N}(\mathrm{d}x; \sigma_z, 1) \\
&= \int f\big[\exp(\sigma_{\phi_0(z)}^2/2 - \sigma_{\phi_0(z)} x)\big] \cdot \exp(-\sigma_{\phi_0(z)}^2/2 + \sigma_{\phi_0(z)} x)\mathcal{N}(\mathrm{d}x; 0, 1) \\
&= \int f\big(w^{-1}\big) w \cdot Q_{\phi_0(z)}(\mathrm{d}w) \\
&= \int f\big(w\big) w^{-1} \cdot Q^{1/\cdot}_{\phi_0(z)}(\mathrm{d}w).
\end{aligned}$$

One can also consider, with $z \mapsto \omega_z = \omega_{\phi_0(z)} > 0$,

$$Q_z(\mathrm{d}w) = \frac{\omega_z}{1 + \omega_z}\delta_{\omega_z}(\mathrm{d}w) + \frac{1}{1 + \omega_z}\delta_{1/\omega_z}(\mathrm{d}w),$$

because

$$\begin{aligned}
\int f(w) Q_z(\mathrm{d}w) &= \frac{\omega_z}{1 + \omega_z} f(\omega_z) + \frac{1}{1 + \omega_z} f(\omega_z^{-1}) \\
&= \omega_z \frac{1}{1 + \omega_z} f(\omega_z) + \omega_z^{-1} \frac{\omega_z}{1 + \omega_z} f(\omega_z^{-1}) \\
&= \int f(w) w Q^{1/\cdot}_{\phi_0(z)}(\mathrm{d}w).
\end{aligned}$$

Example 12 (Reversible jump MCMC Green (1995))**.** Here we are concerned with the situation where X is a disjoint union, for example $\mathsf{X} = \bigsqcup_{i \in \mathbb{N}} \{i\} \times \mathsf{X}_i$ with, for $i \in \mathbb{N}$, $(\mathsf{X}_i, \mathscr{X}_i)$ a measurable space and $\mathscr{X}$ a sigma algebra associated to X; see Fremlin (2010, 214K) for a construction. Here the probability distribution of interest is $\pi(i, \mathrm{d}x_i)$, that is for $i \in \mathbb{N}$, $\pi(i, \cdot)\colon \mathscr{X}_i \mapsto \mathbb{R}_+$ is a finite measure and $\sum_{i=1}^{\infty} \pi(i, \mathsf{X}_i) = 1$. The idea of Green (1995) to circumvent the possibly differing nature of the X_i's is to introduce the following space and probability embeddings:

(a) $E := \bigsqcup_{i,j \in \mathbb{N}} \{(i,j)\} \times \mathsf{X}_i \times \mathsf{U}_{ij}$ such that for $(i,j) \in \mathbb{N}^2$ there exist measurable bijections $\mathsf{X}_i \times \mathsf{U}_{ij} \to \mathsf{X}_j \times \mathsf{U}_{ji}$ for the measurable sapces $\big(\mathsf{X}_i \times \mathsf{U}_{ij}, \mathscr{X}_i \otimes \mathscr{U}_{ij}\big)$ and $\big(\mathsf{X}_j \times \mathsf{U}_{ji}, \mathscr{X}_j \otimes \mathscr{U}_{ji}\big)$;

(b) for $(i,j) \in \mathbb{N}^2$ one chooses mappings $\phi_{ij} = \phi_{ji}^{-1}$ and define $\phi\colon E \to E$ the $\phi(i, j, x_i, u_{ij}) := \big(j, i, \phi_{ij}(x_i, u_{ij})\big)$.

(c) the probability distribution π is embedded in $\mu(i, j, d(x_i, u_{ij})) = \pi(i, \mathrm{d}x_i)\mu_i(j, \mathrm{d}u_{ij} \mid x_i)$.

This can be viewed as a natural generalization of Remark 5.

Remark 9. In light of Example 3 and its relation to the framework in Tierney (1998), it is natural to ask whether the framework considered here is more powerful in terms of its ability to express and validate Markov kernels. In fact it is not, but is perhaps more natural to use since one does not introduce additional auxiliary variables in μ. In particular, for a given choice of μ and ϕ, one can always embed $(\xi, \phi(\xi))$ in the extended space $E \times E$ with distribution $\tilde{\mu}(\mathrm{d}x, \mathrm{d}y) = \mu(\mathrm{d}x)\delta_{\phi(x)}(\mathrm{d}y)$, and use the involution $\tilde{\phi}(x, y) = (y, x)$. The $\tilde{\mu}$-reversibility then follows from Theorem 3. For an expression for the acceptance ratio, it is then convenient to consider the $\tilde{\phi}$-invariant reference measure $\upsilon = \lambda + \lambda^{\tilde{\phi}}$.

We obtain that $\mathrm{d}\tilde{\mu}/\mathrm{d}\upsilon(x) = \rho(x)\left\{1 + \frac{\mathrm{d}\lambda^{\phi}}{\mathrm{d}\lambda}(x)\right\}^{-1}$, where $\rho = \mathrm{d}\mu/\mathrm{d}\lambda$. We obtain that for x in the same $S = S(\mu, \mu^{\phi})$,

$$r((x,\phi(x)),(\phi(x),x)) = \frac{\frac{\mathrm{d}\tilde{\mu}^{\tilde{\phi}}}{\mathrm{d}\upsilon}(x)}{\frac{\mathrm{d}\tilde{\mu}}{\mathrm{d}\upsilon}(x)} = \frac{\rho\circ\phi}{\rho}(x)\frac{1+\frac{\mathrm{d}\lambda^{\phi}}{\mathrm{d}\lambda}(x)}{1+\frac{\mathrm{d}\lambda}{\mathrm{d}\lambda^{\phi}}(x)} = \frac{\rho\circ\phi}{\rho}(x)\frac{\mathrm{d}\lambda^{\phi}}{\mathrm{d}\lambda}(x),$$

as in Proposition 2.

4 Beyond reversibility and standard deterministic proposals

Reversibility plays a central role in the design of MCMC algorithms but is not necessarily a desirable property. In fact, it has been shown that nonreversible Markov chains can converge more quickly in some cases (Diaconis, Holmes, and Neal, 2000), and their ergodic averages can have smaller asymptotic variance in comparison to a suitable reversible counterpart (Neal, 2004; Sun, Schmidhuber, and Gomez, 2010; Chen and Hwang, 2013; Christophe Andrieu, 2016). This can be intuitively attributed to the fact that reversible processes tend to backtrack and/or move in a diffusive way, suggesting slower exploration of the target distribution in comparison to nonreversible processes that move in a more systematic way through the state space.

We discuss here a popular class of nonreversible MH type updates which can be understood as being the cycle of two $\mu-$reversible Markov kernels. This type of non reversibility is referred to as $(\mu, \mathfrak{S})-$reversibility in the literature (Christophe Andrieu and Livingstone, 2019) and was first discussed in Yaglom (1949) as a generalisation of deterministic time-reversible systems. The necessity for some of the conditions below is discussed in Thin, Durmus, et al. (2020).

Proposition 3. *Let μ be a probability distribution on $(E, \mathscr{E})$, $\phi, \sigma \colon E \to E$ be involutions with σ such that $\mu^{\sigma} = \mu$. Let*

(a) Π be the $\mu-$reversible Markov kernel using ϕ and acceptance function $a(r) = 1 \wedge r$,

(b) $\mathfrak{S}$ be such that for $\xi \in E$, $\mathfrak{S}(\xi, \{\sigma(\xi)\}) = 1$ (or for $\xi' \in E$, $\mathfrak{S}(\xi, \mathrm{d}\xi') = \delta_{\sigma(\xi)}(\mathrm{d}\xi')$),

(c) $\lambda \gg \mu$ be such that $\lambda \equiv \lambda^{\phi}$ and $\lambda^{\sigma} = \lambda$.

Let $\psi := \sigma \circ \phi$ and Ψ such that for $\xi \in E$, $\Psi(\xi, \{\psi(\xi)\}) = 1$ (or for $\xi' \in E$, $\Psi(\xi, \mathrm{d}\xi') = \delta_{\psi(\xi)}(\mathrm{d}\xi')$) then

(a) $\psi^{-1} = \sigma \circ \psi \circ \sigma$, $\lambda^{\phi} = \lambda^{\psi^{-1}}$, and $\rho \circ \sigma = \rho$,

(b) the $\mu-$invariant cycle $\mathit{\Pi} := \Pi\mathfrak{S}$ is given by

$$\mathit{\Pi}(\xi, \mathrm{d}\xi') = a \circ r(\xi) \cdot \Psi\big(\xi, \mathrm{d}\xi'\big) + [1 - a \circ r(\xi)]\mathfrak{S}(\xi, \mathrm{d}\xi'),$$

where with $S = \{\xi \in E \colon \rho(\xi) \wedge [\rho \circ \psi(\xi)\mathrm{d}\lambda^{\psi^{-1}}/\mathrm{d}\lambda(\xi)] > 0\}$,

$$r(\xi) = \begin{cases} \frac{\rho\circ\psi}{\rho}(\xi)\frac{\mathrm{d}\lambda^{\psi^{-1}}}{\mathrm{d}\lambda}(\xi) & \xi \in S, \\ 0 & \textit{otherwise.} \end{cases}$$

(c) In fact $\mathit{\Pi}$ is $(\mu, \mathfrak{S})-$reversible (or satisfied the modified or skew detailed balance), that is for $\xi, \xi' \in E$

$$\mu(\mathrm{d}\xi)\mathit{\Pi}(\xi, \mathrm{d}\xi') = \mu(\mathrm{d}\xi')\mathfrak{S}\mathit{\Pi}\mathfrak{S}(\xi', \mathrm{d}\xi).$$

(d) *Let* $\mu(\mathrm{d}\xi) := \pi(\mathrm{d}\xi_0)\mu_{\xi_0}(\mathrm{d}\xi_{-0})$, *where* μ_{ξ_0} *denotes the conditional distribution of* ξ_{-0} *given* ξ_0 *under* μ. *Assume* Π *to be* $(\mu, \mathfrak{S})-$*reversible, where* $\mathfrak{S}(\xi_0, \xi_{-0}) = \big(\mathfrak{S}_0(\xi_0), \xi_{-0}\big)$ *for* $\mathfrak{S}_0$ *and involution. Then the Markov kernel*

$$\Pi_0(\xi_0, A) := \int \mathbf{1}_A(\xi_0')\mu_{\xi_0}(\mathrm{d}\xi_{-0})\Pi(\xi; \mathrm{d}\xi'), \qquad A \in \mathscr{Z},$$

is $(\pi, \mathfrak{S}_0)-$*reversible.*

(e) *Let* $\Pi' := \mathfrak{S}\Pi$, *then with* $\psi' := \phi \circ \sigma$ *and* $\Psi'(\xi, \mathrm{d}\xi') = \delta_{\psi'(\xi)}(\mathrm{d}\xi')$ *then Properties (a)-(d) hold with* Π, ψ, Ψ *and* r *replaced with* Π', ψ', Ψ' *and* $r' := r \circ \sigma$.

Corollary 1. *If* ψ *in Proposition 3 preserves* λ *then one has* $r(\xi) = \rho \circ \psi / \rho\,(\xi)$ *on* $S = \{\xi \in E\colon \rho(\xi) \wedge \rho \circ \psi(\xi) > 0\}$. *Indeed, if* $\lambda^{\psi} = \lambda$ *then* $\lambda^{\psi \circ \psi^{-1}} = \lambda^{\psi^{-1}}$ *so* $\lambda^{\psi^{-1}} = \lambda$ *also.*

Corollary 2. *In many situations, nonreversible kernels are given in the form of* Π *or* Π', *where* $\psi, \psi'\colon E \to E$ *are invertible mappings with the property that* $\psi^{-1} = \sigma \circ \psi \circ \sigma$ *for* σ *an involution leaving* μ *and* λ *invariant, and similarly for* ψ'. *This time-reversal feature ensures that we are in the setup of Proposition 3, since indeed in this setup* $\phi := \sigma \circ \psi$ *(or* $\tilde{\phi} := \psi \circ \sigma$*) is an involution, therefore defining* Π *satisfying the right property. In particular we always have the decomposition* $\Pi = \mathfrak{S}\tilde{\Pi} = \Pi\mathfrak{S}$ *where* Π *and* $\tilde{\Pi}$ *satisfy detailed balance Christophe Andrieu and Livingstone (2019, Theorem 4).*

Remark 10. Proposition 3 highlights the fundamental difference between reversible and this type of nonreversible kernels. Without refreshment of ξ_{-0}, the reversible Markov chain started at ξ oscillates between ξ and $\phi(\xi)$ due to the involutive property, while the nonreversible chain can in principle explore a large subset of states $\psi^k(\xi)$, $k \in \mathbb{N}$, although rejection leads to backtracking. This fundamental qualitative behaviour is exploited in more general and realistic setups, even when ξ_{-0} is refreshed.

Remark 11. In the same way the results of Maire, Douc, and Olsson (2014) can be used in the context of Proposition 1 (see Remark 4) one can, for example, deduce optimality properties of Π_0 from those of Π by using Christophe Andrieu and Livingstone (2019).

In practice, a number of deterministic transformations ψ are used to define π-invariant Markov kernels. The validity of such kernels often rests primarily on showing that the transformation is measure-preserving, typically with the measure being the Lebesgue measure. We give here some examples where π is a probability measure associated with a position variable $x \in \mathbb{R}^d$ and a velocity variable $v \in \mathbb{R}^d$.

A general class of nonreversible MH kernels relies on the choices $\xi = (x, v) \in E = \mathsf{X} \times \mathsf{V}$, $\sigma(x, v) = (x, -v)$ and $\mu(\mathrm{d}x, \mathrm{d}v) = \pi(\mathrm{d}x)\kappa(\mathrm{d}v)$ where κ is such that $\mu^{\sigma} = \mu$. In order to keep presentation simple we will assume that $\mathsf{X} = \mathsf{V} = \mathbb{R}^d$ and that μ has a density $\rho(x, v) = \varpi(x)\kappa(v)$ with respect to the Lebesgue measure on $\mathbb{R}^{2d}$. Note that the Lebesgue measure is invariant by σ since its Jacobian is 1.

Lemma 2. *Let* $x \in \mathsf{X} \subseteq \mathbb{R}^d$ *and* $y \in \mathsf{Y} \subseteq \mathbb{R}^{d'}$, *and* $\psi : \mathsf{X} \times \mathsf{Y} \to \mathsf{X} \times \mathsf{Y}$ *be defined as* $\psi(x, y) = (x, y + f(x))$ *for some function* $f : \mathsf{X} \to \mathsf{Y}$. *Then* ψ *preserves the Lebesgue measure* λ *on* $\mathsf{X} \times \mathsf{Y}$.

Example 13 (Guided Random walk (GRW), Gustafson 1998)**.** Let $\psi(x, v) = (x + v, v)$ then $\phi = \sigma \circ \psi = (x + v, -v)$ is an involution, and is in fact the involution used to define the random walk Metropolis. Then ψ preserves λ by Lemma 2. Hence, using that $\kappa(-v) = \kappa(v)$ for $\xi \in S$,

$$r(x, v) = \frac{\rho \circ \psi}{\rho}(x, v) = \frac{\varpi(x + v)}{\varpi(x)},$$

which coincides with the acceptance ratio of the RWM Metropolis. In fact $P(x, \mathrm{d}x') := \int \kappa(\mathrm{d}v)\Pi(x, v; \mathrm{d}x')$ is the $\pi-$reversible RWM Markov kernel. The GRW, of transition Π, differs in that it is μ-invariant but not reversible and has the property that it introduces memory on the velocity component of the process. On its own Π does not lead to an ergodic chain and must be combined with other updates, e.g. occasionally sampling v afresh from κ.

Before covering Hamiltonian Monte Carlo, and in particular the common variant using the velocity Verlet, or leapfrog, integrator we note that transformations ψ satisfying $\psi^{-1} = \sigma \circ \psi \circ \sigma$ are particularly intuitive in that the iterated maps $\psi \circ \cdots \circ \psi$ can be "reversed".

Remark 12. Let $\psi^0 = \mathrm{Id}$ and $\psi^k = \psi \circ \psi^{k-1}$ for $k \in \mathbb{N}$. If ψ satisfies $\psi^{-1} = \sigma \circ \psi \circ \sigma$, then ψ is time-reversible in the sense that $\phi_k = \sigma \circ \psi^k$ is an involution for any $k \in \mathbb{N}$. Indeed, we have

$$\mathrm{Id} = \psi^{-k} \circ \psi^k = (\sigma \circ \psi \circ \sigma)^k \circ \psi^k = \sigma \circ \psi^k \circ \sigma \circ \psi^k.$$

Lemma 3. *Let $x, v \in \mathbb{R}^d$ and $\psi : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ be*

$$\psi = \psi_B \circ \psi_A \circ \psi_B,$$

where $\psi_B = (x, v) \mapsto (x, v + \imath(x))$ and $\psi_A = (x, v) \mapsto (x + \jmath(v), v)$ for some functions $\imath \colon \mathsf{X} \to \mathsf{V}$ and $\jmath \colon \mathsf{V} \to \mathsf{X}$, where $\jmath(-v) = -\jmath(v)$ for $v \in \mathbb{R}^d$. Then ψ_A, ψ_B and ψ preserve the Lebesgue measure on $\mathbb{R}^{2d}$ and $\psi^{-1} = \sigma \circ \psi \circ \sigma$ so that ψ is time-reversible in the sense of Remark 12.

Example 14 (HMC - leapfrog integrator)**.** Let π have density $\rho = \varpi \otimes \kappa$ w.r.t. λ, the Lebesgue measure on $E = \mathbb{R}^{2d}$. Consider the function $h = \psi_B \circ \psi_A \circ \psi_B$ as in Lemma 3 with $\imath(x) = \frac{\epsilon}{2} \nabla \log \varpi(x)$ and $\jmath(v) = \epsilon \nabla \log \kappa(v)$. Let $\mathit{\Pi} = \Pi\mathfrak{S}$ be the nonreversible kernel in Proposition 3 with $\psi = h^k$ for some $k \in \mathbb{N}$, and acceptance ratio

$$r(\xi) = \frac{\rho \circ \psi}{\rho}(\xi), \qquad \xi \in S.$$

This kernel is a version of the HMC kernel with leapfrog integrator (see Remark 13 below). It has desirable properties, but it is also clear that the μ-invariance of $\mathit{\Pi}$ applies for a much broader class of $\imath$ and $\jmath$, as implied by the appeal to Lemma 3. For example, it is well known that one could replace ϖ in $\imath$ with some approximate density (see, e.g., Neal et al., 2011, Section 5.5), i.e. run the "leapfrog integrator" for a different density but accept or reject using $\rho = \varpi \otimes \kappa$. In order to preserve persistence of motion (and nonreversibility) this update is typically combined with partial refreshment of the velocity. As discussed below, full refreshment leads to a reversible algorithm.

Remark 13. It is often the case, as was the case in part of the seminal paper of Horowitz (1991), that the kernel considered is reversible. Indeed in those works the kernel considered is, for $(x, A) \in \mathsf{X} \times \mathscr{X}$

$$P(x, A) := \int \kappa(\mathrm{d}v) \mathit{\Pi}(x, v; \mathrm{d}(y, w)) \mathbb{I}\{(y, w) \in A \times \mathsf{V}\},$$

that is the velocity is refreshed at each iteration and with $\xi = (x, v)$ and $f, g \colon \mathsf{X} \to [0, 1]$,

$$\begin{aligned}
\int f(x) g(x') \pi(\mathrm{d}x) P(x, \mathrm{d}x') &= \int f(x) g(x') \mu(\mathrm{d}\xi) \mathit{\Pi}(\xi, \mathrm{d}\xi') \\
&= \int f(x) g(x') \mu(\mathrm{d}\xi') \mathfrak{S} \mathit{\Pi} \mathfrak{S}(\xi, \mathrm{d}\xi') \\
&= \int f(x) g(x') \mu \mathfrak{S}(\mathrm{d}\xi') \mathfrak{S} \mathit{\Pi}(\xi, \mathrm{d}\xi') \\
&= \int f(x) g(x') \mu(\mathrm{d}\xi') \mathit{\Pi}(\xi, \mathrm{d}\xi') \\
&= \int f(x) g(x') \pi(\mathrm{d}x') P(x', \mathrm{d}x).
\end{aligned}$$

where we have used the $(\mu, \mathfrak{S})-$reversibility of $\mathit{\Pi}$, $\mu = \mu\mathfrak{S}$ and the fact that $\mathfrak{S} g = g$ for this choice of function.

Example 15 (MALA and generalized MALA). Standard, reversible normal (i.e. κ is the standard normal distribution) MALA (Besag, 1994) corresponds to one iteration of HMC - leapfrog integrator with full refreshment of the velocity at each iteration, and indeed here $\psi(x,v) = \big(x + \imath(x) + \frac{\epsilon}{2}v, -v - \frac{\epsilon}{2}\big[\imath(x) + \imath\big(x + \frac{\epsilon}{2}\imath(x) + \frac{\epsilon}{2}v\big)\big]\big)$ for $\imath(x) = \nabla_x \log \varpi(x)$ and $\epsilon > 0$. In Poncet (2017) it is proposed to consider $\imath\colon \mathsf{X}\times\{-1,1\} \to \mathsf{X}$ with $\imath(x,s) = \nabla_x \log \varpi(x) + s\gamma(x)$ for $\gamma\colon \mathsf{X} \to \mathsf{X}$. A naïve idea would be to take $\psi_B = (x,v,s) \mapsto (x, v + \imath(x,s), s)$ and $\psi_A = (x,v,s) \mapsto (x + \jmath(v), v, s)$ with $\sigma(x,v,s) = (x,v,-s)$ which is shown to have poor properties; this leads to the development of a scheme relying on an implicit integration scheme.

Example 16 (Hyperplane reflection). If λ_V is the Lebesgue measure on $\mathbb{R}^d$, the involution $\mathrm{b}(x,v) = (x, v - 2\{n(x)^\top v\}n(x))$ preserves $\lambda = \lambda_X \times \lambda_V$, where $n : \mathbb{R}^d \to \mathbb{R}^d$ satisfies $\|n(x)\|^2 = n(x)^\top n(x) = 1$ for all $x \in \mathbb{R}^d$. Indeed, we can write the v-component of $\phi(x,v)$ as

$$v - 2(n(x)^\top v)n(x) = (\mathrm{Id} - 2n(x)n(x)^\top)v,$$

and we see that $\mathrm{B}_x = (\mathrm{Id} - 2n(x)n(x)^\top)$ is a matrix with $\mathrm{B}_x^2 = \mathrm{Id}$ and so $|\det \mathrm{B}_x| = 1$. Since b does not move the x-component, it follows that b is λ-preserving.

If $\lambda = \lambda_X \times \lambda_V$ and λ_V is instead the uniform measure on the sphere $\mathbb{S}^{d-1} = \{v \in \mathbb{R}^d : v^\top v = 1\}$ then $\mathrm{B}_x^\top = \mathrm{B}_x$ so $\|\mathrm{B}_x v\|^2 = (\mathrm{B}_x v)^\top \mathrm{B}_x v = v^\top v = \|v\|^2$, so B_x preserves the norm $\|\cdot\|$. Letting Leb denote the Lebesgue measure on $\mathbb{R}^d$, and noting from the argument above that b preserves Leb, we can then conclude that b is λ-preserving as above because for any measurable $A \subseteq \mathbb{S}^{d-1}$, $\lambda_V(A) \propto \mathrm{Leb}(\{tv : v \in A, t \in [0,1]\})$.

A natural question is whether the requirement that $\sigma\colon \mathsf{E} \to \mathsf{E}$ be an involution can be relaxed to invertibility only. More precisely let $\psi_0, \sigma_0\colon \mathsf{Z} \to \mathsf{Z}$ be invertible with $\psi_0^{-1} = \sigma_0^{-1} \circ \psi_0 \circ \sigma_0$–such a structure is known as time-reversible symmetry when ψ_0 is the flow of a dynamical system with this property (Lamb and J. A. Roberts, 1998). Let $\Psi_0(z, \mathrm{d}z') = \delta_{\psi_0(z)}(\mathrm{d}z')$, $\mathfrak{S}_0^{\pm 1}(z, \mathrm{d}z') = \delta_{\sigma_0^{\pm}(z)}(\mathrm{d}z')$ and μ_0 a probability distribution on $\big(\mathsf{Z}, \mathscr{Z}\big)$ such that $\mu_0\mathfrak{S}_0^{\pm 1} = \mu_0$. Can one define a deterministic MH type kernel leaving μ_0 invariant – Fang, J.-M. Sanz-Serna, and Skeel (2014) provide us with an answer, see below. Our answer consists of embedding this problem in the $(\mu, \mathfrak{S})-$reversible framework. Let $\mathsf{E} = \mathsf{Z} \times \mathsf{U}$ where $\mathsf{U} = \{-1, 1\}$ and define $\mu\big(\mathrm{d}(z,u)\big) := \frac{1}{2}\mu_0(\mathrm{d}z)\mathbb{I}\big\{u \in \mathsf{U}\big\}$ and consider the mappings $\sigma, \psi\colon \mathsf{Z} \times \mathsf{U} \to \mathsf{Z} \times \mathsf{U}$ such that for $f\colon \mathsf{Z} \times \mathsf{U} \to \mathbb{R}$, $f \circ \sigma(z,u) = f\big(\sigma_0^u(z), -u\big)$ and $\psi(z,u) := \big(\sigma_0^{-u} \circ \psi_0^u(z), u\big)$. For any $(z,u) \in \mathsf{Z} \times \mathsf{U}$ we have that $\sigma^2(z,u) = \big(\sigma_0^{-u} \circ \sigma_0^u(z), u\big) = (z,u)$, that is σ is an involution and one can check that $\psi^{-1}(z,u) = \big(\psi_0^{-u} \circ \sigma_0^u(z), u\big)$. Noting that $\psi_0^{-1} = \sigma_0^{-1} \circ \psi_0 \circ \sigma_0$ is equivalent to $\psi_0^{-u} = \sigma_0^{-u} \circ \psi_0^u \circ \sigma_0^u$ for $u \in \mathsf{U}$ we have for $(z,u) \in \mathsf{Z} \times \mathsf{U}$

$$\begin{aligned}
\sigma \circ \psi \circ \sigma(z,u) &= \sigma \circ \psi\big(\sigma_0^u(z), -u\big) \\
&= \sigma\big(\sigma_0^u \circ \psi_0^{-u} \circ \sigma_0^u(z), -u\big) \\
&= \big(\psi_0^{-u} \circ \sigma_0^u(z), u\big) \\
&= \psi^{-1}(z,u).
\end{aligned}$$

Finally, for any $f\colon \mathsf{Z}\times\mathsf{U}\to\mathbb{R}$ we have $\mu\mathfrak{S}=\mu$, since

$$\begin{aligned}\int f(z,u)\mu^{\sigma}\big(\mathrm{d}(z,u)\big) &= \int f\circ\sigma(z,u)\mu\big(\mathrm{d}(z,u)\big)\\ &= \int f\big(\sigma_0^u(z),-u\big)\frac{1}{2}\mu_0(\mathrm{d}z)\mathbb{I}\big\{u\in\mathsf{U}\big\}\\ &= \int f\big(z,-u\big)\frac{1}{2}\mu_0^{\sigma_0^u}(\mathrm{d}z)\mathbb{I}\big\{u\in\mathsf{U}\big\}\\ &= \int f\big(z,u\big)\frac{1}{2}\mu_0(\mathrm{d}z)\mathbb{I}\big\{u\in\mathsf{U}\big\}\\ &= \int f(z,u)\mu\big(\mathrm{d}(z,u)\big),\end{aligned}$$

We are therefore back in the $(\mu,\mathfrak{S})-$reversible setup and with

$$\begin{aligned}\alpha(z,u) &:= a\left(\frac{\rho\circ\psi(z,u)}{\rho(z,u)}\frac{\mathrm{d}\lambda^{\psi^{-1}}}{\mathrm{d}\lambda}(z,u)\right)\\ &= a\left(\frac{\rho_0\circ\sigma_0^{-u}\circ\psi_0^u(z)}{\rho_0(z)}\frac{\mathrm{d}\lambda^{\psi_0^{-u}\circ\sigma_0^u}}{\mathrm{d}\lambda_0}(z)\right)\\ &= a\left(\frac{\rho_0\circ\psi_0^u(z)}{\rho_0(z)}\frac{\mathrm{d}\lambda^{\psi_0^{-u}}}{\mathrm{d}\lambda_0}(z)\right)\end{aligned}$$

we can define the kernel

$$\begin{aligned}\mathit{\Pi} f(z,u) &= \alpha(z,u)\cdot f\circ\psi(z,u)+\bar{\alpha}(z,u)\cdot f\circ\sigma(z,u)\\ &= \alpha(z,u)\cdot f\big(\sigma_0^{-u}\circ\psi_0^u(z),u\big)+\bar{\alpha}(z,u)\cdot f\big(\sigma_0^u(z),-u\big).\end{aligned}$$

The kernel $\widetilde{\mathit{\Pi}}\colon \mathsf{Z}\times\mathscr{Z}\to[0,1]$ proposed by Fang, J.-M. Sanz-Serna, and Skeel (2014) is, for $g\colon \mathsf{Z}\to\mathbb{R}$,

$$\begin{aligned}\widetilde{\mathit{\Pi}} g(z) &= \alpha(z,1)g\circ\psi_0(z)+\bar{\alpha}(z,1)g\circ\sigma_0(z)\\ &= \alpha(z,1)\tilde{g}\circ\sigma\circ\psi(z,1)+\bar{\alpha}(z,1)\tilde{g}\circ\sigma(z,1)\end{aligned}$$

where $\tilde{g}\colon \mathsf{Z}\times\mathsf{U}\to\mathbb{R}$ is such that $\tilde{g}(z,1)=\tilde{g}(z,-1):=g(z)$, which can therefore be thought of as being $\mathit{\Pi}$ but used for the value $u=1$ only–one could equally have chosen $u=-1$, naturally. One can check that this kernel satisfies global balance for μ_0 directly (Fang, J.-M. Sanz-Serna, and Skeel, 2014). The kernel $\widetilde{\mathit{\Pi}}$ does not satisfy detailed or skew detailed balance, but noting that $\phi:=\sigma\circ\psi$ is an involution and letting $\Phi(z,\mathrm{d}z'):=\delta_{\phi(z)}(\mathrm{d}z')$, that is $\Phi f(z,u)=(\psi_0^u,-u)$ we use "reversibility" (see the proof Proposition 3) to show

$$\begin{aligned}\frac{1}{2}\int\alpha(z,1)g\circ\psi_0(z)\mu_0(\mathrm{d}z) &= \int\mathbf{1}_{\mathsf{Z}\times\{1\}}(z,u)\tilde{g}\circ\phi(z,u)\alpha(z,u)\mu(\mathrm{d}z,u)\\ &= \int\mathbf{1}_{\mathsf{Z}\times\{1\}}(z,u)\Phi\tilde{g}(z,u)\alpha(z,u)\mu(\mathrm{d}z,u)\\ &= \int\tilde{g}(z,u)\Phi\mathbf{1}_{\mathsf{Z}\times\{1\}}(z,u)\alpha(z,u)\mu(\mathrm{d}z,u)\\ &= \int g(z)\mathbf{1}_{\mathsf{Z}\times\{-1\}}(z,u)\alpha(z,u)\mu(\mathrm{d}z,u)\\ &= \frac{1}{2}\int g(z)\alpha(z,-1)\mu_0(\mathrm{d}z)\end{aligned}$$

and similarly with Φ replaced with $\mathfrak{S}$ (see the proof Proposition 3)

$$\begin{aligned}\frac{1}{2}\int g\circ\sigma_0(z)\alpha(z,1)\mu_0(\mathrm{d}z) &= \int \mathbf{1}_{\mathsf{Z}\times\{1\}}(z,u)\mathfrak{S}\tilde{g}(z,u)\alpha(z,u)\mu(\mathrm{d}z,u)\\ &= \int \tilde{g}(z,u)\mathfrak{S}\mathbf{1}_{\mathsf{Z}\times\{1\}}(z,u)\alpha(z,u)\mu(\mathrm{d}z,u)\\ &= \int g(z)\mathbf{1}_{\mathsf{Z}\times\{-1\}}(z,u)\alpha(z,u)\mu(\mathrm{d}z,u)\\ &= \frac{1}{2}\int g(z)\alpha(z,-1)\mu_0(\mathrm{d}z).\end{aligned}$$

Therefore for any $g\colon \mathsf{Z}\to[0,1]$

$$\int \widetilde{\Pi} g(z)\mu_0(\mathrm{d}z) = \int g\circ\sigma_0(z)\mu_0(\mathrm{d}z) = \int g(z)\mu_0(\mathrm{d}z)$$

and we conclude.

5 Markov chain proposals, stopping times and processes & NUTS

In some scenarios it is desirable for μ to involve simulation of a stopped process. In particular, this allows the amount of simulation required to produce a suitable proposal to be random and ideally be appropriately adapted to features of the target distribution and the current point. As mentioned in Remark 3, the specification of (μ, ϕ) is not unique for a given Markov kernel, so there is some flexibility in precisely how stopping times and stopped processes are captured in ξ and described by μ. In particular, one often has flexibility in allowing ξ to be infinite-dimensional and to contain a realization of the original process as well as the stopping time, or for ξ to be finite-dimensional and to contain only the stopped process. In the former case, one will need to adopt an indirect implementation as per Remark 6.

5.1 A toy example

We illustrate the former approach on a simple example with i.i.d. proposals. Let $\xi = (n, \mathsf{Z}, k) \in \mathbb{N} \times \mathsf{Z}^{\mathbb{N}} \times \mathbb{N}$. Assume that $\nu \gg \pi$ and let $\varpi(z) = \mathrm{d}\pi/\mathrm{d}\nu(z)$ – a common situation is when π and ν have densities w.r.t. some common dominating measure and $\varpi(z) = \pi(z)/\nu(z)$ if we keep the same notation for these densities. Assume that the distribution of Z under μ is that $z_0 \sim \pi$ and for $i \in \mathbb{N}_*$, $z_i \overset{\text{iid}}{\sim} \nu$. Let $(s_n)_{n\in\mathbb{N}_*}$ be a sequence of functions such that $s_n : \mathsf{Z}^{\mathbb{N}} \to \{0, 1\}$ depends only on the first $n + 1$ members of its argument; i.e. $s_n(z)$ depends only on $z_0, \ldots, z_n$. Define the stopping time for $\mathsf{Z} \in \mathsf{Z}^{\mathbb{N}}$

$$\tau = \tau(\mathsf{Z}) := \inf\{n \geq 1 : s_n(\mathsf{Z}) = 1\}. \tag{10}$$

For example, one could choose $s_n(\mathsf{Z}) = \mathbb{I}\{\sum_{i=0}^n \varpi(z_i) > c\}$ for some constant $c > 0$, or $s_n(\mathsf{Z}) = \mathbb{I}\{\mathrm{ESS}(z_0, \ldots, z_n) > c\}$ with

$$\mathrm{ESS}(z_0, \ldots, z_n) := \frac{\{\sum_{i=0}^n \varpi(z_i)\}^2}{\sum_{i=0}^n \varpi(z_i)^2},$$

heuristically to ensure that sufficiently many samples have been drawn and that one can be chosen to produce a sample approximately drawn from π. For $\mathsf{Z} \in \mathsf{Z}^{\mathbb{N}}$, let $n := \tau(\mathsf{Z})$ and $k \sim \varsigma(\cdot; n, \mathsf{Z})$ where $\varsigma(\cdot; n, \mathsf{Z})$ is an arbitrary categorical distribution taking values in $[\![0, n]\!]$ and with probabilities depending on $z_0, \ldots, z_n$ only. For $k \in \mathbb{N}$, let $\sigma_k : \mathsf{Z}^{\mathbb{N}} \to \mathsf{Z}^{\mathbb{N}}$ be the swapping function such that, with $\mathsf{Z}' = \sigma_k(\mathsf{Z})$, $z'_0 = z_k$, $z'_k = z_0$ and $z'_j = z_j$ for $j \notin \{0, k\}$. Clearly, σ_k is an involution and we consider $\phi(n, \mathsf{Z}, k) := (\tau \circ \sigma_k(\mathsf{Z}), \sigma_k(\mathsf{Z}), k)$, which is an involution since

$$\phi \circ \phi(n, \mathsf{Z}, k) = (\tau \circ \sigma_k \circ \sigma_k(\mathsf{Z}), \sigma_k \circ \sigma_k(\mathsf{Z}), k) = (n, \mathsf{Z}, k).$$

Letting $\nu^{\otimes\infty}$ denote the probability measure associated with an infinite sequence of independent ν-distributed random variables, we have for $n \in \mathbb{N}$ and $k \in [\![0, n]\!]$,

$$\mu(n, \mathrm{d}\mathsf{Z}, k) = \varpi(z_0)\nu^{\otimes\infty}(\mathrm{d}\mathsf{Z})\varsigma(k; n, \mathsf{Z})s_n(\mathsf{Z}) \prod_{i=1}^{n-1} \{1 - s_i(\mathsf{Z})\},$$

where we note that z_0 is marginally distributed according to π, which together with ϕ above defines the kernel outlined in Alg. 2. One can check that the acceptance ratio is, with $(n', \mathsf{Z}', k') = \phi(n, \mathsf{Z}, k) = (\tau \circ \sigma_k(\mathsf{Z}), \sigma_k(\mathsf{Z}), k)$ and for $\xi \in S = \{(n, \mathsf{Z}, k) : n = \tau(\mathsf{Z}), 0 \leq k \leq \tau(\mathsf{Z}) \wedge \tau \circ \sigma_k(\mathsf{Z})\}$,

$$r(\xi) = \frac{\varpi(z'_0)\varsigma(k'; n', \mathsf{Z}')}{\varpi(z_0)\varsigma(k; n, \mathsf{Z})} = \frac{\varpi(z_k)\varsigma(k; n', \mathsf{Z}')}{\varpi(z_0)\varsigma(k; n, \mathsf{Z})}.$$

Although theoretically convenient, the algorithm described in Alg. 2 is not very practical due to the requirement to sample the infinite-dimensional Z_{-0}. However the definitions of $(s_n)_{n\in\mathbb{N}_*}$, $\tau(\mathsf{Z})$, k and

Algorithm 2 Impractical algorithm

(a) Simulate (lazily) $z_i \overset{\text{iid}}{\sim} \nu$, for $i \in \mathbb{N}_*$.

(b) Set $n \leftarrow \tau(\mathtt{Z})$.

(c) Simulate $k \sim \varsigma(\cdot \mid n, \mathtt{Z})$.

(d) Set $\mathtt{Z}' \leftarrow \sigma_k(\mathtt{Z})$ and $n' \leftarrow \tau \circ \sigma_k(\mathtt{Z})$.

(e) With probability

$$a\left(\frac{\varpi(z_k)\varsigma(k; n', \mathtt{Z}')}{\varpi(z_0)\varsigma(k; n, \mathtt{Z})}\right),$$

output z_k, otherwise output z_0.

$\sigma_k(\mathtt{Z})$ are such that $\mathtt{Z}$ is not required in its entirety to simulate from the kernel, which can be achieved with finite computation provided $\tau(\mathtt{Z}) < \infty$. This is described in Alg. 3, with a slight abuse of notation since s_k and σ_k are defined on $\mathtt{Z}^{\mathbb{N}}$. We will refer to this as a "lazy" implementation or simulation and adopt the presentation in Alg. 2 for brevity. In particular, the explicit lazy implementation in Alg. 3 involves simulating only those components of $\mathtt{Z}$ and $\mathtt{Z}'$ that are required to implement Alg. 2, the details of which are fairly straightforward and tend to obscure the simplicity of the approach. Note that throughout we give the expression for the acceptance ratio on S only in order to alleviate presentation.

If we choose $(s_n)_{n\in\mathbb{N}_*}$ such that $\tau(\mathtt{Z}) = 1$ and $\varsigma(1; 1, \mathtt{Z}) = 1$ for all $\mathtt{Z} \in \mathtt{Z}^{\mathbb{N}}$ then this reduces to the independent MH (IMH), but of course in general it allows more than one candidate sample from ν to be simulated. We refer to the kernel in Alg. 2 as an adaptive IMH kernel for this reason. If we let $\varsigma(k; n, \mathtt{Z}) \propto \varpi(z_k)\mathbf{1}_{[\![0,n]\!]}(k)$ then we obtain

$$r(\xi) = \frac{\sum_{i=0}^{n} \varpi(z_i)}{\sum_{i=0}^{n'} \varpi(z'_i)} \mathbf{1}_{[\![0,n\wedge n']\!]}(k)$$

An important point is that n' may not equal n in general, requiring in particular additional simulations when $n' > n$. By choosing $(s_n)_{n\in\mathbb{N}}$ and $\varsigma(\cdot; n, \mathtt{Z})$ appropriately one can ensure that $n = n'$ for all $\mathtt{Z} \in \mathtt{Z}^{\mathbb{N}}$. This has the appeal that there is no need to perform additional simulations once $z_1, \ldots, z_\tau$ are realized, and can also mean that the acceptance ratio is one The following lemma provides sufficient conditions for this equality to hold.

Lemma 4. *Let $\mathtt{Z} \in \mathtt{Z}^{\mathbb{N}}$ be such that, with the definition in (10), $\tau(\mathtt{Z}) < \infty$ and assume further that $(s_k)_{k\in\mathbb{N}_*}$ satisfies*

(a) $k \mapsto s_k(\mathtt{Z})$ is non-decreasing;

(b) $s_k(\mathtt{Z}) = s_k \circ \sigma_l(\mathtt{Z})$ for all $l \in [\![k]\!]$;

Then $\tau \circ \sigma_l(\mathtt{Z}) = \tau(\mathtt{Z})$ for $l \in [\![\tau(\mathtt{Z}) - 1]\!]$.

Example 17. For $k \in \mathbb{N}$, let $s_k(\mathtt{Z}) := \mathbb{I}\{\sum_{i=0}^{k} \varpi(z_i) > c\}$ for some constant $c > 0$ and let $\mathtt{Z} \in \mathtt{Z}^{\mathbb{N}}$ satisfy $\tau(\mathtt{Z}) < \infty$. Clearly $k \mapsto s_k(\mathtt{Z})$ is non-decreasing and condition (a) in Lemma 4 holds. Assume that for $n \in \mathbb{N}$, $s_1(\mathtt{Z}) = \cdots = s_{n-1}(\mathtt{Z}) = 0$ while $s_n(\mathtt{Z}) = 1$. It follows that for $l \in [\![0, n-1]\!]$, if $k \in [\![n-1]\!]$ then $s_k \circ \sigma_l(\mathtt{Z}) \leq \mathbb{I}\{\sum_{i=0}^{n-1} \varpi(z_i) > c\} = s_{n-1}(\mathtt{Z}) = s_k(\mathtt{Z}) = 0$ while if $k \geq n$ then $s_k \circ \sigma_l(\mathtt{Z}) \geq \mathbb{I}\{\sum_{i=0}^{n} \varpi(z_i) > c\} = s_n(\mathtt{Z}) = 1 = s_k(\mathtt{Z})$ so condition (b) in Lemma 4 holds. If we take

Algorithm 3 Practical, lazy implementation

(a) Set $i \leftarrow 1$, simulate $z_1 \sim \nu$.

(b) While $s_i(z_{0:i}) = 0$

(i) Set $i \leftarrow i + 1$.

(ii) Simulate $z_n \sim \nu$.

(c) Set $n \leftarrow i = \tau(\mathsf{Z})$.

(d) Simulate $k \sim \varsigma(\cdot; n, \mathsf{Z})$ and set $z'_{0:n} = \sigma_k(z_{0:n})$.

(e) Set $i \leftarrow 1$.

(f) While $s_i(z'_{0:i}) = 0$

(i) Set $i \leftarrow i + 1$.

(ii) If $i > n$, simulate $z'_i = z_i \sim \nu$.

(g) Set $n' \leftarrow i = \tau \circ \sigma_k(\mathsf{Z}) = \tau(\mathsf{Z}')$.

(h) With probability

$$a\left(\frac{\varpi(z_k)\varsigma(k; n', \mathsf{Z}')}{\varpi(z_0)\varsigma(k; n, \mathsf{Z})}\right),$$

output z_k, otherwise output z_0.

$\varsigma(k; n, \mathsf{Z}) \propto \varpi(z_k)\mathbf{1}_{[\![0,n-1]\!]}(k)$ then the conclusion of Lemma 4 holds and we obtain for $k \in [\![0, n-1]\!]$, with $\xi = (n, \mathsf{Z}, k)$ and $\phi(n, \mathsf{Z}, k) := (\tau \circ \sigma_k(\mathsf{Z}), \sigma_k(\mathsf{Z}), k)$,

$$r(\xi) = \frac{\sum_{i=0}^{n-1} \varpi(z_i)}{\sum_{i=0}^{n-1} \varpi(z'_i)} = \frac{\sum_{i=0}^{n-1} \varpi(z_i)}{\sum_{i=0}^{n-1} \varpi(z_i)} = 1.$$

In contrast, for the choice $s_n(\mathsf{Z}) = \mathbb{I}\{\mathrm{ESS}(z_0, \ldots, z_n) > c\}$, condition (a) in Lemma 4 is not satisfied and there is no reason for $n = n'$ to hold for all $\mathsf{Z} \in \mathsf{Z}^{\mathbb{N}}$. We expand on the idea of Lemma 4 to validate the NUTS and stopping-time MTM in Sections 5 and 6.

5.2 Doubly-infinite Markov chain proposal and change of measure

Here we demonstrate how one can verify that Markov chain proposals can be used to define π-invariant Markov kernels. First we show how to deal with a distribution μ involving a doubly-infinite Markov chain as well as a proposal index. Then we consider the more involved but practical scenario where the proposal index is selected from a window of random size, adapted according to user-defined constraint functions. A special case of this framework, and indeed the inspiration for the generalization here, is when the Markov chain is a deterministic dynamical system is the No U-Turn Sampler (NUTS) of Hoffman and Gelman (2014).

Let π and ν be measures on $(\mathsf{Z}, \mathscr{Z})$ where π is a probability, $\nu \gg \pi$ and let $\varpi := \mathrm{d}\pi/\mathrm{d}\nu$. In order to present our algorithm we require the definition of a two sided Markov chain, from which a proposal state is chosen within a MH kernel update.

Definition 4 (Two-sided (k, π, Q, Q^*)-Markov chain probability measure Λ^k)**.** Let π be a probability measure on $(\mathsf{Z}, \mathscr{Z})$ and Q, $Q^* \colon \mathsf{Z} \times \mathscr{Z} \to [0, 1]$ be transition kernels. Then for $k \in \mathbb{Z}$, denote by Λ^k

the probability measure on $\left(\mathsf{Z}^{\mathbb{Z}}, \mathscr{Z}^{\otimes\mathbb{Z}}\right)$ associated with the Markov chain Z such that $Z_k \sim \pi$, and for $i \in \mathbb{N}_*$, $Z_{i+k} \mid \{Z_{i+k-1} = z\} \sim Q(z, \cdot)$ and $Z_{k-i} \mid \{Z_{k-i+1} = z\} \sim Q^*(z, \cdot)$.

We define $\mathcal{Q}(z_0, \cdot)$ to be the probability measure for $\mathtt{Z} \sim \Lambda^0$ conditional on a fixed z_0.

For any $\mathtt{Z} \in \mathsf{Z}^{\mathbb{Z}}$ we let $\varsigma(\cdot; \mathtt{Z})$ be a probability distribution on $\left(\mathbb{Z}, \mathscr{P}(\mathbb{Z})\right)$ and we are interested in the update outlined in Alg. 4, where for $i \in \mathbb{Z}$, $\theta^i : \mathsf{Z}^{\mathbb{Z}} \to \mathsf{Z}^{\mathbb{Z}}$ is the shift function defined via $\theta^i(\mathtt{Z})_j = z_{i+j}$ and to ease the presentation of the algorithms, we write that one should "lazily" simulate a realization of a double-infinite Markov chain, by which we mean that only a finite number of states of the Markov chain should be required to perform the rest of the algorithm. The simulation of $\mathtt{Z}$ is naturally not practical and a stopping criterion is required, while making sense of the acceptance ratio and its expression require an additional assumption on (ν, Q, Q^*). These are the topics of the remainder of the subsection.

Algorithm 4 To sample from $P_{\mathrm{MC}}^{(\mathrm{general})}(z_0, \cdot)$

(a) Lazily simulate $\mathtt{Z} \sim \mathcal{Q}(z_0, \cdot)$.

(b) Simulate $k \sim \varsigma(\cdot; \mathtt{Z})$.

(c) With probability

$$a\left(\frac{\varpi(z_k)\varsigma\big(-k; \theta^k(\mathtt{Z})\big)}{\varpi(z_0)\varsigma(k; \mathtt{Z})}\right),$$

output z_k. Otherwise output z_0.

We first introduce an assumption on (ν, Q, Q^*) justifying the form of the acceptance ratio in full generality.

Definition 5 (Reversible triplet (ν, Q, Q^*))**.** Let ν be a measure on $(\mathsf{Z}, \mathscr{Z})$ and $Q, Q^* \colon \mathsf{Z} \times \mathscr{Z} \to [0, 1]$ be two Markov kernels. We say that (ν, Q, Q^*) is a reversible triplet if for $z, z' \in \mathsf{Z}$,

$$\nu(\mathrm{d}z)Q(z, \mathrm{d}z') = \nu(\mathrm{d}z')Q^*(z', \mathrm{d}z), \tag{11}$$

This implies in particular that Q is ν-invariant, whether ν is a probability measure or not, and Q^* is the time-reversal of Q. In operator theoretic language Q^* is the $\nu-$adjoint of Q for the inner product $\langle f, g \rangle = \int f g \mathrm{d}\nu$ on $L^2(\nu)$. Importantly for practical purposes, we observe that (11) accommodates invertible mappings that leave ν invariant:

Proposition 4. *Let ν be a measure on $(\mathsf{Z}, \mathscr{Z})$ and $\psi \colon \mathsf{Z} \to \mathsf{Z}$ be invertible and such that $\nu^\psi = \nu$. Then (11) holds with $Q(z, \mathrm{d}z') = \Psi(z, \mathrm{d}z') := \delta_{\psi(z)}(\mathrm{d}z')$ and $Q^*(z, \mathrm{d}z') = \Psi^*(z, \mathrm{d}z') := \delta_{\psi^{-1}(z)}(\mathrm{d}z')$.*

The assumption that (ν, Q, Q^*) is a reversible triplet implies that for any $k \in \mathbb{Z}$, Λ^0 and Λ^k are equivalent on a suitable restriction of $\mathsf{Z}^{\mathbb{Z}}$, with a simple Radon–Nikodym derivative involving ϖ only. This is the property used in Alg. 4 to propose that a chain $\mathtt{Z}$ distributed according to Λ^0 is mapped to a chain distributed according to Λ^k.

Lemma 5. *Let π and ν be measures on $(\mathsf{Z}, \mathscr{Z})$ where π is a probability and $\nu \gg \pi$ and let $\varpi := \mathrm{d}\pi/\mathrm{d}\nu$. Assume that (ν, Q, Q^*) is a reversible triplet. For any $k \in \mathbb{Z}$ let Λ^k be the two-sided (k, π, Q, Q^*)-Markov chain probability measure and $S_k := \{\mathtt{Z} \in \mathsf{Z}^{\mathbb{Z}} : \varpi(z_0) \wedge \varpi(z_k) > 0\}$. Then for any $k \in \mathbb{Z}$ and $\mathtt{Z} \in S_k$,*

$$\frac{\mathrm{d}\Lambda^k_{S_k}}{\mathrm{d}\Lambda^0_{S_k}}(\mathtt{Z}) = \frac{\varpi(z_k)}{\varpi(z_0)}.$$

We can now establish correctness of Alg. 4.

Corollary 3. *For any* $\mathtt{Z} \in \mathsf{Z}^{\mathbb{Z}}$ *let* $\varsigma(\cdot;\mathtt{Z})$ *be a probability distribution on* $\big(\mathbb{Z}, \mathscr{P}(\mathbb{Z})\big)$, $\xi := (\mathtt{Z}, k) \in \mathsf{Z}^{\mathbb{Z}} \times \mathbb{Z}$, $\mu(\mathrm{d}\mathtt{Z}, k) = \Lambda^0(\mathrm{d}\mathtt{Z})\varsigma(k;\mathtt{Z})$ *and define the involution* $\phi(\mathtt{Z}, k) := (\theta^k(\mathtt{Z}), -k)$. *Then for* $\xi \in S = \big\{(\mathtt{Z}, k) : \varpi(z_0) \wedge \varpi(z_k) \wedge \varsigma(k;\mathtt{Z}) \wedge \varsigma\big(-k;\theta^k(\mathtt{Z})\big) > 0\big\}$, *we have*

$$\frac{\mathrm{d}\mu_S^\phi}{\mathrm{d}\mu_S}(\xi) = \frac{\varpi(z_k)}{\varpi(z_0)} \frac{\varsigma\big(-k;\theta^k(\mathtt{Z})\big)}{\varsigma(k;\mathtt{Z})},$$

and apply Theorem 3.

Alg. 4 is in general not practical due to the requirement of simulation from $\mathcal{Q}(z_0, \cdot)$, a prerequisite to sample from $\varsigma(\cdot;\mathtt{Z})$. Key to this is to make the dependence of $\varsigma(\cdot;\mathtt{Z})$ on $\mathtt{Z} \in \mathsf{Z}^{\mathbb{Z}}$ "finite", that is dependent on a finite number of coordinates of $\mathtt{Z}$ in order to ensure a finite amount of computation. Numerous options are possible and we outline two here. The first one is purely deterministic.

Example 18. Let $\tau \in \mathbb{N}$ and assume that for any $\mathtt{Z} \in \mathsf{Z}^{\mathbb{Z}}$ the probability $\varsigma(\cdot;\mathtt{Z})$ is entirely determined by the $2\tau+1$ states $z_{-\tau}, \ldots, z_\tau$ and of support $[\![-\tau, \tau]\!]$. In this case simulating $k \sim \varsigma(\cdot;\mathtt{Z})$ only requires simulation of this subsequence. However in order to compute the acceptance ratio it is required to simulate what is unrealized in the subsequence $z_{k-\tau}, \ldots, z_{k+\tau}$, that is z_i for $i \in [\![k-\tau, k+\tau]\!] \cap [\![-\tau, \tau]\!]^{\complement}$. An example for $\varsigma(\cdot;\mathtt{Z})$ is $\varsigma(k;\mathtt{Z}) \propto \mathbf{1}_{[\![-\tau,\tau]\!]}(k)\varpi(z_k)$, in which case the acceptance ratio is $\sum_{i=k-\tau}^{k+\tau} \varpi(z_i) / \sum_{i=-\tau}^{\tau} \varpi(z_i)$ on S.

The above example, in the context of HMC, gives a simple version of what is described in Neal (1994), which can of course be embellished in various ways. It is also possible to adapt τ to the realization $\mathtt{Z}$.

Example 19. It is possible to make τ a function $\mathtt{Z} \mapsto \tau(\mathtt{Z}) \in \mathbb{N}$ in Example 18, more precisely a stopping time adapted to sequences of the form $z_{-i}, z_{-i+1}, \ldots, z_0, \ldots, z_{i-1}, z_i$ such that with $n = \tau(\mathtt{Z})$, sampling $k \sim \varsigma(\cdot;\mathtt{Z})$ is entirely determined by $z_{-n}, \ldots, z_0, \ldots, z_n$. This leads to the same need for additional simulation i.e. z_i for $i \in [\![k - \tau \circ \theta^k(\mathtt{Z}), k + \tau \circ \theta^k(\mathtt{Z})]\!] \cap [\![-\tau(\mathtt{Z}), \tau(\mathtt{Z})]\!]^{\complement}$ where we notice the need to determine the value of the stopping time value for the sequence $\theta^k(\mathtt{Z})$, also required for the computation of the acceptance ratio $\sum_{i=k-\tau\circ\theta^k(\mathtt{Z})}^{k+\tau\circ\theta^k(\mathtt{Z})} \varpi(z_i) / \sum_{i=-\tau(\mathtt{Z})}^{\tau(\mathtt{Z})} \varpi(z_i)$ on S, for the choice $\varsigma(k;\mathtt{Z}) \propto \mathbf{1}_{[\![-\tau(\mathtt{Z}),\tau(\mathtt{Z})]\!]}(k)\varpi(z_k)$.

In the next section we explore a general technique of ensuring that both windows of states coincide, therefore leading to simplified algorithms.

Remark 14. One could considerably weaken the condition (11) in Lemma 5 to

$$\nu(\mathrm{d}z)Q(z, \mathrm{d}z') = \nu^*(\mathrm{d}z')Q^*(z', \mathrm{d}z), \tag{12}$$

where ν and ν^* are equivalent but not necessarily equal, at the expense of simplicity. In this case, we obtain for $\mathtt{Z} \in S = \{\mathtt{Z} : \varpi(z_0) \wedge \varpi(z_k) > 0\}$,

$$\frac{\mathrm{d}\Lambda_S^k}{\mathrm{d}\Lambda_S^0}(\mathtt{Z}) = \frac{\varpi(z_k)}{\varpi(z_0)} F_k(\mathtt{Z}),$$

where

$$F_k(\mathtt{Z}) = \begin{cases} \prod_{i=1}^{k} \frac{\mathrm{d}\nu}{\mathrm{d}\nu^*}(z_i) & k > 0, \\ \prod_{i=k+1}^{0} \frac{\mathrm{d}\nu^*}{\mathrm{d}\nu}(z_i) & k < 0, \\ 1 & k = 0. \end{cases}$$

When $\nu = \nu^\psi$, that is ψ is $\nu-$preserving, then $\mathrm{d}\nu^\psi/\mathrm{d}\nu = 1$. The generality of (12) is natural in the context of deterministic, invertible maps ψ that are not measure preserving but such that $\nu^\psi \equiv \nu$. In particular, the analogue of Proposition 4 holds with $\nu^* = \nu^\psi$.

Lemma 6. *Let $\psi\colon \mathsf{Z} \to \mathsf{Z}$ be invertible and such that $\nu^{\psi} \equiv \nu$. Then with $\Psi(z, \mathrm{d}z') = \delta_{\psi(z)}(\mathrm{d}z')$ and $\Psi^*(z, \mathrm{d}z') = \delta_{\psi^{-1}(z)}(\mathrm{d}z')$ then with $\nu^* = \nu^{\psi}$,*

$$\nu(\mathrm{d}z)\Psi(z, \mathrm{d}z') = \nu^*(\mathrm{d}z')\Psi^*(z', \mathrm{d}z).$$

In the specific case that ν is the Lebesgue measure and ψ^{-1} is a diffeomorphism, then

$$\frac{\mathrm{d}\nu^{\psi}}{\mathrm{d}\nu}(z) = \left|\det(\psi^{-1})'(z)\right|$$

is the Jacobian.

5.3 Doubly-infinite Markov chain proposal and coinciding windows

In Examples 18 and 19, the two windows around z_0 and z_k are typically different when $k \neq 0$. We now explain how to devise an instance of the framework where the windows around z_0 and z_k are identical by construction. The main idea consists of introducing an auxiliary variable ℓ that can be thought of as determining the left index of the realized window. To be precise, let $m \in \mathbb{N}$ be the fixed size of the window to be realized and $\xi := (\mathsf{Z}, \ell, k) \in \mathsf{Z}^{\mathbb{Z}} \times \mathbb{N} \times \mathbb{N}$ where $\ell \sim \mathrm{Uniform}(\llbracket 0, m-1 \rrbracket)$ and $k \sim \varsigma(k; \ell, \mathsf{Z}) = \varsigma(k; \ell, z_{-\ell}, \ldots, z_r)$ with $r := m - 1 - \ell$, that is here $\mu(\mathrm{d}\xi) := \Lambda_0(\mathrm{d}\mathsf{Z})\varsigma(k; \ell, \mathsf{Z})\mathbb{I}\{\ell \in \llbracket 0, m-1 \rrbracket\}/m$. Now define the involution

$$\phi(\xi) = (\theta^k(\mathsf{Z}), \ell + k, -k),$$

then, observing that by construction $\theta^k(\mathsf{Z})_{-(\ell+k):(r-k)} = (z_{-\ell}, \ldots, z_r)$, we obtain the acceptance ratio

$$r(\mathsf{Z}, \ell, k) = \frac{\varpi(z_k)\varsigma(-k; \ell + k, z_{-\ell}, \ldots, z_r)}{\varpi(z_0)\varsigma(k; \ell, z_{-\ell}, \ldots, z_r)},$$

on $S = \{\xi\colon \varpi(z_k) \wedge \varpi(z_0) \wedge \varsigma(-k; \ell + k, z_{-\ell}, \ldots, z_r) \wedge \varsigma(k; \ell, z_{-\ell}, \ldots, z_r) > 0\}$. For example, if $\varsigma(k; \ell, z_{-\ell}, \ldots, z_r) \propto \mathbf{1}_{\llbracket -\ell, r \rrbracket}(k)\varpi(z_k)$ then the acceptance ratio is 1 for all $k \in \llbracket -\ell, r \rrbracket$ such that $\varsigma(k; \ell, z_{-\ell}, \ldots, z_r) > 0$. The resulting algorithm is presented in Alg. 5.

Algorithm 5 To sample from $P_{\mathrm{MC}}^{(\mathrm{general})}(z_0, \cdot)$

(a) Lazily simulate $\mathsf{Z} \sim \mathcal{Q}(z_0, \cdot)$.

(b) Simulate $\ell \sim \mathrm{Uniform}(\llbracket 0, m-1 \rrbracket)$ and $k \sim \varsigma(k; \ell, z_{-\ell}, \ldots, z_r)$.

(c) With probability

$$a\left(\frac{\varpi(z_k)\varsigma(-k; \ell + k, z_{-\ell}, \ldots, z_r)}{\varpi(z_0)\varsigma(k; \ell, z_{-\ell}, \ldots, z_r)}\right),$$

output z_k. Otherwise output z_0.

In order to introduce NUTS-like kernels, it is helpful at this point to consider the case where $m = 2^n$ for some $n \in \mathbb{N}$ and we shall reparameterize ℓ as a sequence of n bits. That is, we define $b \in \{0,1\}^n$ to be a sequence of independent Bernoulli(1/2) random variates and write $\ell = \sum_{j=1}^n b_j 2^{j-1}$, so that the distribution of ℓ is indeed $\mathrm{Uniform}(\llbracket 0, m-1 \rrbracket)$. In order to specify the involution in this reparameterization we define $\beta : \llbracket 0, 2^{n-1} \rrbracket \to \{0,1\}^n$ to be the function that computes the "reversed" binary representation of its input with n bits, e.g. $\beta(13) = (1,0,1,1,0)$, which has the property that $\beta \circ \ell(b) = b$. Finally, we specify $\phi(\mathsf{Z}, b, k) = (\theta^k(\mathsf{Z}), \beta(\ell(b) + k), -k)$. The intuition is that given Z, the binary string b defines a particular window $z_{-\ell}, \ldots, z_r$ around z_0 and for a given $k \in \llbracket -\ell, r \rrbracket$, there is a corresponding binary string $\beta\big(\ell(b) + k\big)$ that defines the same window, but around z_k.

Algorithm 6 To sample from $P_{\mathrm{MC}}^{(\mathrm{NUTS})}(z_0, \cdot)$

NUTS-like algorithm

(a) Lazily simulate $\mathtt{Z} \sim \mathcal{Q}(z_0, \cdot)$ and b.

(b) Sample (ℓ, r):

(i) Set $n \leftarrow 0$ and $\ell_0 \leftarrow r_0 \leftarrow 0$.

(ii) Set $n \leftarrow n + 1$.

(iii) Set $\ell_n \leftarrow \ell_{n-1} + b_n 2^{n-1}$ and $r_n \leftarrow r_{n-1} + (1 - b_n)2^{n-1}$.

(iv) If $s_n(\mathtt{Z}, b) = 0$, go to 2(b)

(v) Set $\ell \leftarrow \ell_{n-1}$, $r \leftarrow r_{n-1}$.

(c) Sample $k \sim \varsigma(\cdot \mid \mathtt{Z}, \ell, r)$.

(d) With probability

$$a\left(\frac{\varpi(z_k)\varsigma(-k \mid \theta^k(\mathtt{Z}), \ell + k, r - k)}{\varpi(z_0)\varsigma(k \mid \mathtt{Z}, \ell, r)}\right),$$

output z_k. Otherwise output z_0.

5.4 NUTS-like kernels

Let π and ν be measures on $(\mathsf{Z}, \mathscr{Z})$ where π is a probability and $\nu \gg \pi$ and let $\varpi := \mathrm{d}\pi/\mathrm{d}\nu$.

Define for $n \in \mathbb{N}$,

$$\ell_n(b) := \sum_{j=1}^{n} b_j 2^{j-1}, \qquad r_n(b) := 2^n - 1 - \ell_n(b).$$

Let $(s_n)_{n\in\mathbb{N}}$ be a sequence of functions where $s_n \colon \mathsf{Z}^{\mathbb{Z}} \times \{0,1\}^{\mathbb{N}} \to \{0,1\}$ depends only on windows of states in a way that is made clear below. For $(\mathtt{Z}, b) \in \mathsf{Z}^{\mathbb{N}} \times \{0,1\}^{\mathbb{N}}$ define the stopping time

$$\tau(b, \mathtt{Z}) := \inf\{n \geq 1 : s_n(\mathtt{Z}, b) = 1\}$$

Specifically, we require that $s_n(\mathtt{Z}, b)$ is a function of the vector $(z_{-\ell_n(b)}, \ldots, z_{r_n(b)})$.

Note that $\ell_n(b)$ and $r_n(b)$ can be computed recursively, which suggests step (b) in Alg. 6 where for a sequence random variables $b = (b_1, b_2, \ldots) \in \{0,1\}^{\mathbb{N}}$, $b_i \overset{\text{iid}}{\sim} \text{Bernoulli}(1/2)$ one finds $\tau := \tau(b, \mathtt{Z})$ and the final window is defined by $\ell = \ell_{\tau-1}(b)$ and $r = r_{\tau-1}(b)$, i.e. the most recently added states are ignored. The reason for this will become clearer below, but is essentially analogous to the argument in Example 17. We now turn to the specification of $(s_n)_{n\in\mathbb{N}}$. For $(n, b) \in \mathbb{N} \times \{0,1\}^{\mathbb{N}}$ define $m_n(b) := 2^{n-1} - 1 - \ell_n(b)$, so that $[\![-\ell_n(b), m_n(b)]\!]$ and $[\![m_n(b) + 1, r_n(b)]\!]$ are integer sequences of length 2^{n-1}, one of which is $[\![\ell_{n-1}(b), r_{n-1}(b)]\!]$:

- if $b_n = 0$ then $\ell_{n-1}(b) = \ell_n(b)$ and $r_{n-1}(b) = r_n(b) - 2^{n-1} = 2^n - 1 - \ell_n(b) - 2^{n-1} = m_n(b)$,
- if $b_n = 1$ then $\ell_{n-1}(b) = \ell_n(b) - 2^{n-1} = -(m_n(b) + 1)$ and $r_{n-1}(b) = r_n(b)$

Let for $n \in \mathbb{N}_*$

$$s_n(\mathtt{Z}, b) = f_{n-1}(z_{-\ell_n(b)}, \ldots, z_{m_n(b)}) \vee f_{n-1}(z_{m_n(b)+1}, \ldots, z_{r_n(b)}),$$

where $\left\{f_k \colon \mathsf{Z}^{2^k} \to \{0,1\}, k \in \mathbb{N}\right\}$ is defined recursively via

$$f_k(\mathfrak{z}_{1:2^k}) = \begin{cases} g_k(\mathfrak{z}_{1:2^k}) \vee f_{k-1}(\mathfrak{z}_{1:2^{k-1}}) \vee f_{k-1}(\mathfrak{z}_{(2^{k-1}+1):2^k}) & k \in \mathbb{N}_* \\ g_0(\mathfrak{z}_1) & k = 0 \end{cases},$$

for some functions $\left\{g_k : \mathsf{Z}^{2^k} \to \{0,1\}, k \in \mathbb{N}\right\}$, which encode the condition for stopping while the functions $f_{k-1}(\mathfrak{z}_{1:2^{k-1}})$ and $f_{k-1}(\mathfrak{z}_{(2^{k-1}+1):2^k})$ report whether stopping was triggered in either of the two main subtrees.

Example 20 (HMC-NUTS)**.** Assume the setup of Example 14 that is with $z = (x, v) \in \mathsf{X} \times \mathsf{V}$ we target $\pi(x, v) \propto \gamma(x)\kappa(v)$ and let $Q(z, \mathrm{d}z') = \Psi(z, \mathrm{d}z') = \delta_{\psi(z)}(\mathrm{d}z')$ for ψ the leapfrog mapping and assume that the dominating measure satisfies $\nu(\mathrm{d}z)\Psi(z, \mathrm{d}z') = \nu(\mathrm{d}z')\Psi^*(z', \mathrm{d}z)$ with $\Psi^*(z, \mathrm{d}z') = \delta_{\psi^{-1}(z)}(\mathrm{d}z')$. A possible choice is for $k \in \mathbb{N}_*$ and $\ell, r \in \mathbb{Z}^2$ such that $\ell - r + 1 = 2^k$,

$$g_k(z_\ell \ldots, z_r) = \mathbb{I}\left\{(x_r - x_\ell)^\top v_\ell > 0\right\} \vee \mathbb{I}\left\{(x_r - x_\ell)^\top v_r < 0\right\} \vee \mathbb{I}\left\{\max_{i,j \in [\![\ell, r]\!]} \pi(z_i)/\pi(z_j) > \Delta_{\max}\right\}, \tag{13}$$

for some $\Delta_{\max} > 0$, and $g_0 \equiv 0$. The first two indicators correspond to the choice made in Hoffman and Gelman (2014), the motivation for NUTS–see Appendix D for some details. The last indicator is our own suggestion to address numerical errors, since the setup considered by Hoffman and Gelman (2014) corresponds to Example (21) below and numerical errors are addressed in a slightly different way.

Definition 6 (Slice sampler Besag et al. (1995) and Neal (2003))**.** Given a target distribution π, with density ϖ w.r.t. some dominating measure ν, one can define an extended target distribution $\tilde{\pi}$ via the decomposition

$$\varpi(z, u) = \varpi(z)\mathbb{I}\{u \leq \varpi(z)\}/\varpi(z) = \mathbb{I}\{u \leq \varpi(z)\},$$

where ϖ is the density of $\tilde{\pi}$ w.r.t. the product of ν and the Lebesgue measure, and in which conditional on z, u is uniformly distributed on $[0, \varpi(z)]$. A MwG Markov kernel leaving $\tilde{\pi}$ invariant consists of sampling u uniformly on $[0, \varpi(z)]$ and then applying any Markov kernel leaving the conditional distribution π_u of z given u invariant: the uniform distribution on the "slice" $\{z \in \mathsf{Z} : \varpi(z) \geq u\}$. For the purposes of this work, we may seek to define a sophisticated π_u-invariant Markov kernel.

Example 21 (sliced-HMC-NUTS)**.** This is what Hoffman and Gelman (2014) refer to as the "simplified" NUTS algorithm. Here the overall target distribution has density $\eta(x, v, u) \propto \mathbb{I}\{u \leq \gamma(x)\kappa(v)\}$ and the algorithm consists of a MwG alternating between updating u given $z = (x, v)$ and vice versa. Given $u \in [0, \gamma \otimes \kappa(z)]$, we focus on sampling from $\pi_u(x, v) \propto \mathbb{I}\{u \leq \gamma(x)\kappa(v)\}$. In this scenario, in addition to (13) it is suggested to use, for $k \in \mathbb{N}_*$ and $\ell, r \in \mathbb{Z}^2$ such that $\ell - r + 1 = 2^k$,

$$g_k(z_\ell \ldots, z_r) = \mathbb{I}\left\{(x_r - x_\ell)^\top v_\ell > 0\right\} \vee \mathbb{I}\left\{(x_r - x_\ell)^\top v_r < 0\right\},$$

and

$$g_0(z) = \mathbb{I}\left\{\log \gamma \otimes \kappa(z) < \log u - \Delta_{\max}\right\},$$

for $\Delta_{\max} \geq 0$ in order to stop computation when the error arising from the numerical integration of Hamilton's dynamic leads to an "astronomically" large error.

Lemma 7. *The functions* $\{s_n, n \in \mathbb{N}\}$ *satisfy for any* $(n, \mathsf{Z}, b) \in \mathbb{N} \times \mathsf{Z}^{\mathbb{Z}} \times \{0,1\}^{\mathbb{N}_*}$

(a) $s_n(\mathsf{Z}, b)$ *depends only on the order and the values of* $z_{-\ell_n(b)}, \ldots, z_{r_n(b)}$*, and not on how they are indexed;*

(b) $s_n(\mathsf{Z}, b) \geq s_{n-1}(\mathsf{Z}, b)$ *for* $n \geq 2$*,*

(c) $s_n(\mathsf{Z}, b) \geq g_{n-1}(z_{-\ell_{n-1}(b)}, \ldots, z_{r_{n-1}(b)})$*.*

Remark 15. Part (c) of Lemma 7 is useful for computational reasons: if one observes that $s_n(\mathtt{Z}, b) = 0$ but $g_n(z_{-\ell_n(b)}, \dots, z_{r_n(b)}) = 1$ then one can stop and take $\tau(\mathtt{Z}, b) = n + 1$ without further simulation being needed.

Let $\beta_k \colon [\![0, 2^k - 1]\!] \to \{0,1\}^k$ be the reversed binary representation of $i \in [\![0, 2^k - 1]\!]$ with k bits, so that e.g., $\beta_5(13) = (1, 0, 1, 1, 0)$. This function has the property that $\beta_k \circ \ell_k(b) = b_{1:k}$. For $b \in \{0,1\}^{\mathbb{N}}$, define $\bar{b}_i := (b_{i+1}, b_{i+2}, \dots)$ for $i \in \mathbb{N}$. We can now present the result that allows one to relate the stopped Markov processes, which is analogous to Lemma 4 in this setting.

Lemma 8. *Let $n \in \mathbb{N}$, $\ell, r \in \mathbb{N}_*$ be such that $\ell + r + 1 = 2^{n-1}$, $k \in [\![-\ell, r]\!]$, $\mathtt{Z} \in \mathtt{Z}^{\mathbb{Z}}$ and $\mathtt{Z}' := \theta^k(\mathtt{Z})$. Then for any $b = (b_{1:n-1}, \bar{b}_{n-1}) \in \{0,1\}^{\mathbb{N}_*}$ such that $\tau(\mathtt{Z}, b) = n$, $\ell_{n-1}(b) = \ell$ and $r_{n-1}(b) = r$, there exists a unique $b'_{1:n-1} = \beta_{n-1}(\ell + k)$ such that with $b' = (b'_{1:n-1}, \bar{b}_{n-1})$,*

(a) $\ell_{n-1}(b') = \ell + k$,

(b) $r_{n-1}(b') = r - k$,

(c) $(z'_{-(\ell+k)}, \dots, z'_{r-k}) = (z_{-\ell}, \dots, z_r)$,

(d) and $\tau(\mathtt{Z}', b') = n$.

Corollary 4. *For $n \in \mathbb{N}_*$ and $k \in [\![-\ell_{n-1}(b), r_{n-1}(b)]\!]$ let $\chi_{n-1,k} \colon \{0,1\}^{\mathbb{N}_*} \to \{0,1\}^{\mathbb{N}_*}$ such that $\chi_{n-1,k}(b) := \big(\beta_{n-1}(\ell_{n-1}(b) + k), \bar{b}\big)$. Then $(b, k) \mapsto (\chi_{n-1,k}(b), -k)$ is an involution since $b'_{1:n-1} = \beta_{n-1}\big(\ell_{n-1}(b) + k\big)$, $\ell_{n-1}(b') = \ell_{n-1}(b) + k$ and $b_{1:n-1} = \beta_{n-1}\big(\ell_{n-1}(b) + k - k\big)$.*

For $n \in \mathbb{N}_*$ let $\mathfrak{s}(n, b) \colon \mathtt{Z}^{\mathbb{Z}} \times \mathbb{N} \times \{0,1\}^{\mathbb{N}_*} \to \{0,1\}$

$$\mathfrak{s}(\mathtt{Z}, n, b) := s_n(\mathtt{Z}, b) \prod_{i=1}^{n-1} \left[1 - s_i(\mathtt{Z}, b)\right],$$

$\xi := (\mathtt{Z}, n, b, k) \in \mathtt{Z}^{\mathbb{Z}} \times \mathbb{N} \times \{0,1\}^{\mathbb{N}} \times \mathbb{Z}$ and

$$\mu(\mathrm{d}\mathtt{Z}, n, b, k) := \Lambda^0(\mathrm{d}\mathtt{Z})\Upsilon(b)\mathfrak{s}(\mathtt{Z}, n, b)\varsigma\big(k; \ell_{n-1}(b), r_{n-1}(b), \mathtt{Z}\big).$$

From Lemma 8 for $(n, b) \in \mathbb{N} \times \{0,1\}^{\mathbb{N}}$ $\mathfrak{s}(n, b) = 1$ implies that $\mathfrak{s}(n, \chi_{n-1,k}(b)) = 1$ for any $k \in [\![-\ell_{n-1}(b), r_{n-1}(b)]\!]$. Then with the involution

$$\phi(\mathtt{Z}, n, b, k) = (\theta^k(\mathtt{Z}), n, \chi_{n-1,k}(b), -k)$$

we define

$$\begin{aligned} S := \{(\mathtt{Z}, n, b, k) \in \Xi : \varpi(z_k) \wedge \varpi(z_0) \wedge \mathfrak{s}(\mathtt{Z}, n, b) \wedge \mathfrak{s}\big(\mathtt{Z}, n, \chi_{n-1,k}(b)\big) \\ \wedge \varsigma(k; \ell_{n-1}(b), r_{n-1}(b), \mathtt{Z}) \wedge \varsigma(-k; \ell_{n-1}(b) + k, r_{n-1}(b) - k, \theta^k(\mathtt{Z})) > 0\}, \end{aligned}$$

Now for any $\xi \in S$ we obtain an expression for the acceptance ratio

$$\begin{aligned} r(\xi) &= \frac{\mathrm{d}\Lambda^k}{\mathrm{d}\Lambda^0}(\mathtt{Z}) \frac{\varsigma(-k; \theta^k(\mathtt{Z}), \ell + k, r - k)}{\varsigma(k; \mathtt{Z}, \ell, r)} \\ &= \frac{\varpi(z_k)}{\varpi(z_0)} \frac{\varsigma(-k; \theta^k(\mathtt{Z}), \ell + k, r - k)}{\varsigma(k; \mathtt{Z}, \ell, r)}. \end{aligned}$$

A natural choice of $\varsigma(k; \mathtt{Z}, \ell, r)$ is $\varsigma(k; \mathtt{Z}, \ell, r) \propto \varpi(z_k)\mathbb{I}\{k \in [\![-\ell, r]\!]\}$, in which case $r(\mathtt{Z}, \ell, r, k) = 1$ for any $k \in [\![-\ell, r]\!]$ (This is the same in the slice setting, where it corresponds to choosing uniformly from

points in the slice). One can always improve this slightly (Peskun) by excluding $k = 0$, and having an acceptance ratio that is not 1 in general. That is, taking

$$\varsigma(k; \mathsf{Z}, \ell, r) \propto \begin{cases} \varpi(z_k)\mathbb{I}\{k \in [\![-\ell, r]\!] \setminus \{0\}) & \exists i \in [\![-\ell, r]\!] \setminus \{0\} : \varpi(z_i) > 0, \\ \mathbb{I}\{k = 0\} & \text{otherwise.} \end{cases}$$

in which case $r(\xi) = 1$ for $\xi \in S$. Our understanding from Betancourt (2017), is that Stan uses this (unsliced) "multinomial" (i.e. categorical) sampling, but the exact expression for $\varsigma(k; \mathsf{Z}, \ell, r)$ is not clear.

6 Multiple-try Metropolis and related schemes

6.1 Standard MTM

A simple multiple-try Metropolis (MTM) kernel (Liu, Liang, and Wong, 2000) involves $n \in \mathbb{N}$ proposals conditional upon the input, from which one is chosen as a candidate to move to. The acceptance probability then involves simulating $n - 1$ proposals from this candidate. The kernel is presented in Alg. 7. We can write $\xi = (\mathsf{Z}_1, \mathsf{Z}_2, k, \ell)$ where $\mathsf{Z}_1, \mathsf{Z}_2 \in \mathsf{Z}^n$, $k, \ell \in [\![1, n]\!]$, and here $\xi_0 = z_{1,k}$. For simplicity we will assume that the target π and proposals $Q(z, \cdot)$, $z \in \mathsf{Z}$, have densities ϖ and $q(z, \cdot)$ w.r.t. a common reference measure. We can write,

$$\rho(\mathsf{Z}_1, \mathsf{Z}_2, k, \ell) = \frac{\mathbb{I}\{k \in [\![1, n]\!]\}}{n} \varpi(z_{1,k}) \left\{ \prod_{i=1}^{n} q(z_{1,k}, z_{2,i}) \right\} \varsigma(\ell; z_{1,k}, \mathsf{Z}_2) \left\{ \prod_{i=1, i \neq k}^{n} q(z_{2,\ell}, z_{1,i}) \right\}.$$

The associated involution is $\phi(\mathsf{Z}_1, \mathsf{Z}_2, k, \ell) := (\mathsf{Z}_2, \mathsf{Z}_1, \ell, k)$, so that the acceptance ratio is, for $\xi \in S$

$$r(\xi) = \frac{\rho \circ \phi}{\rho}(\xi) = \frac{\varpi(z_{2,\ell}) q(z_{2,\ell}, z_{1,k}) \varsigma(k; z_{2,\ell}, \mathsf{Z}_1)}{\varpi(z_{1,k}) q(z_{1,k}, z_{2,\ell}) \varsigma(\ell; z_{1,k}, \mathsf{Z}_2)}.$$

In practice, one often chooses $\varsigma(\ell; z_{1,k}, \mathsf{Z}_2) \propto w(z_{2,\ell}, z_{1,k})$ where w is a weight function. In particular, Liu, Liang, and Wong (2000) suggest to use

$$w(z, z') = \varpi(z) q(z, z') \lambda(z, z'),$$

where $\lambda(z, z') = \lambda(z', z)$ for all $z, z' \in \mathsf{Z}$. Then the acceptance ratio can be expressed as

$$r(\xi) = \frac{\sum_{i=1}^{n} w(z_{2,i}, z_{1,k})}{\sum_{i=1}^{n} w(z_{1,i}, z_{2,\ell})},$$

for $\xi \in S$. To illustrate, a possible choice is to take $\lambda(z, z') = \{q(z, z') q(z', z)\}^{-1}$, in which case $w(z, z') = \varpi(z)/q(z', z)$.

6.2 Stopping time MTM

We consider now locally adaptive selection of the number of samples n in MTM. In particular, the approach taken in Section 6.1 needs to be slightly adapted and then the stopping time random variables introduced, one for each of the $\mathsf{Z}_1 := (z_{1,1}, z_{1,2}, \ldots) \in \mathsf{Z}^{\mathbb{N}_*}$ samples and the $\mathsf{Z}_2 := (z_{2,1}, z_{2,2}, \ldots) \in \mathsf{Z}^{\mathbb{N}_*}$ samples and for $(m, n) \in \{1, 2\} \times \mathbb{N}_*$ let $\mathsf{Z}_{m,n} := (z_{m,1}, \ldots, z_{m,n})$. The kernel is presented in Alg. 8. We define $\xi := (\mathsf{Z}_1, \mathsf{Z}_2, m, n, k, \ell)$ where $m, n \in \mathbb{N}$, $(\mathsf{Z}_1, \mathsf{Z}_2) \in \mathsf{Z}^{\mathbb{N}_*} \times \mathsf{Z}^{\mathbb{N}_*}$ and $(k, \ell) \in [\![m]\!] \times [\![n]\!]$. We let $\xi_0 = z_{1,1}$. Let $\sigma_k : \mathsf{Z}^{\mathbb{N}_*} \to \mathsf{Z}^{\mathbb{N}_*}$ be the swapping function such that, with $\mathsf{Z}' = \sigma_k(\mathsf{Z})$, $z'_1 = z_k$,

Algorithm 7 Standard MTM kernel

(a) Given z, sample $k \sim \text{Uniform}\big(\llbracket 1, n \rrbracket\big)$ and set $z_{1,k} = z$

(b) Sample $z_{2,i} \overset{\text{iid}}{\sim} Q(z_{1,k}, \cdot)$ for $i \in \llbracket n \rrbracket$

(c) Sample $\ell \sim \varsigma(\cdot; z_{1,k}, \mathsf{Z}_2)$

(d) Sample $z_{1,i} \overset{\text{iid}}{\sim} Q(z_{2,\ell}, \cdot)$ for $i \in \llbracket n \rrbracket \setminus \{k\}$

(e) With probability

$$a\left(\frac{\varpi(z_{2,\ell})q(z_{2,\ell}, z_{1,k})\varsigma(k; z_{2,\ell}, \mathsf{Z}_1)}{\varpi(z_{1,k})q(z_{1,k}, z_{2,\ell})\varsigma(\ell; z_{1,k}, \mathsf{Z}_2)}\right)$$

return $z_{2,\ell}$, otherwise $z_{1,k}$.

$z'_k = z_1$ and $z'_j = z_j$ for $j \notin \{1, k\}$. For any $i \in \mathbb{N}_*$ let $s_i : \mathsf{Z} \times \mathsf{Z}^{\mathbb{N}_*} \to \{0, 1\}$ be such that $i \mapsto s_i(z, \mathsf{Z}')$ is non-decreasing and $s_i(z, \mathsf{Z}') = s_i\big(z, \sigma_\ell(\mathsf{Z}')\big)$ for any $\ell \in \llbracket i - 1 \rrbracket$. For example, one could choose $s_i(z, \mathsf{Z}') = \mathbb{I}\left\{\sum_{j=1}^{i} w(z, z'_i) \geq c\right\}$ for all $i \in \mathbb{N}_*$ and for some $c > 0$, where w is a weight function as described in the previous subsection.

The "forward" stopping times of interest are, for $z \in \mathsf{Z}$ and $\mathsf{Z}_1, \mathsf{Z}_2 \in \mathsf{Z}^{\mathbb{N}}$

$$\tau_1(z, \mathsf{Z}_1) = \inf\{n \geq 1 : s_n(z, \mathsf{Z}_1) = 1\} \text{ and } \tau_2(z, \mathsf{Z}_2) = \inf\{n \geq 1 : s_n(z, \mathsf{Z}_2) = 1\},$$

and we define the $\{0, 1\}$-valued functions, for $n \in \mathbb{N}_*$,

$$\mathfrak{s}(n, z, \mathsf{Z}) = s_n(z, \mathsf{Z}) \prod_{i=1}^{n-1} \big[1 - s_i(z, \mathsf{Z})\big].$$

For $i \in \{1, 2\}$ the quantity $\mathfrak{s}(n, z_{3-i,1}, \mathsf{Z}_i)$ can be thought of as the probability that $\tau_i = n$ given the values $z_{3-i,1}$ and $\mathsf{Z}_{i,n}$. We define $\mathcal{Q}_2(z, \mathrm{d}\mathsf{Z})$ to correspond to the distribution of $z_i \overset{\text{iid}}{\sim} Q(z, \cdot)$ for $i \in \mathbb{N}_*$ and $\mathcal{Q}_1(z, \mathrm{d}\mathsf{Z})$ such that for $i \geq 2$ $z_i \overset{\text{iid}}{\sim} Q(z, \cdot)$ and $\mathcal{Q}_1(z, \mathrm{d}z_1) = \pi(\mathrm{d}z_1)$, that is under μ below $z_{1,1} \sim \pi$,

$$\mu\big(\mathrm{d}(\mathsf{Z}_1, \mathsf{Z}_2), m, n, k, \ell\big) := \mathcal{Q}_2(z_{1,1}, \mathrm{d}\mathsf{Z}_2)\mathfrak{s}(n, z_{1,1}, \mathsf{Z}_2)$$
$$\varsigma(\ell; z_{1,1}, \mathsf{Z}_2, n-1)\mathcal{Q}_1(z_{2,\ell}, \mathrm{d}\mathsf{Z}_1)\mathfrak{s}(m, z_{2,\ell}, \mathsf{Z}_1)\frac{\mathbb{I}\big\{k \in \llbracket m-1 \rrbracket\big\}}{m-1},$$

which is the distribution of a process that simulates the stopped processes described above, and chooses ℓ and k, respectively, from a categorical distribution on $\llbracket n-1 \rrbracket$ and a uniform distribution on $\llbracket m-1 \rrbracket$. We define $\phi(\mathsf{Z}_1, \mathsf{Z}_2, m, n, k, \ell) = (\sigma_\ell(\mathsf{Z}_2), \sigma_k(\mathsf{Z}_1), n, m, \ell, k)$. It is straightforward to verify that ϕ is an involution. What is more interesting is, assuming densities as in the previous subsection, that for $\xi \in S$

$$r(\xi) = \frac{\varpi(z_{2,\ell})q(z_{2,\ell}, z_{1,k})\varsigma(k; z_{2,\ell}, \sigma_k(\mathsf{Z}_1), m-1)}{\varpi(z_{1,1})q(z_{1,k}, z_{2,\ell})\varsigma(\ell; z_{1,1}, \mathsf{Z}_2, n-1)},$$

one can apply Lemma 4 (reindexing to take into account 1-indexing as opposed to 0-indexing) twice to determine that $\mathfrak{s}(n, z_{1,1}, \mathsf{Z}_2) = 1$ implies $\mathfrak{s}\big(n, z_{1,k}, \sigma_\ell(\mathsf{Z}_2)\big) = 1$ and $\mathfrak{s}(m, z_{2,\ell}, \mathsf{Z}_1) = 1$ implies $\mathfrak{s}\big(m, z_{2,\ell}, \sigma_k(\mathsf{Z}_1)\big) = 1$.

Remark 16. It is possible, of course, to specify functions s_i that do not satisfy the conditions above. In this case, the reverse and the forward stopping time probabilities are not necessarily equal, and their ratios will appear in the acceptance ratio.

Algorithm 8 Locally adaptive MTM kernel

(a) Given z set $z_{1,1} = z$

(b) Sample $z_{2,i} \overset{\text{iid}}{\sim} Q(z_{1,1}, \cdot)$ for $i \geq 1$ lazily and obtain $n = \tau_2(z_{1,1}, \mathsf{Z}_2)$

(c) Sample $\ell \sim \varsigma(\cdot; z_{1,1}, \mathsf{Z}_2, n)$

(d) Sample $z_{1,i} \overset{\text{iid}}{\sim} Q(z_{2,\ell}, \cdot)$ for $i \geq 2$ lazily and obtain $m = \tau_1(z_{2,\ell}, \mathsf{Z}_1)$

(e) Sample $k \sim \text{Uniform}(\llbracket m-1 \rrbracket)$.

(f) With probability

$$a\left(\frac{\varpi(z_{2,\ell})q(z_{2,\ell}, z_{1,k})\varsigma(k; z_{2,\ell}, \sigma_k(\mathsf{Z}_1), m-1)}{\varpi(z_{1,1})q(z_{1,k}, z_{2,\ell})\varsigma(\ell; z_{1,1}, \mathsf{Z}_2, n-1)}\right)$$

return $z_{2,\ell}$, otherwise $z_{1,1}$.

6.3 Pseudo-marginal algorithms

It is relatively straightforward to adapt the MTM kernels above to the pseudo-marginal setting (Beaumont, 2003; Christophe Andrieu and G. O. Roberts, 2009). It is also possible to extend the example below to the situation where one uses stopping times to determine the number of simulations, and also to the particle MCMC (Christophe Andrieu, Arnaud Doucet, and Holenstein, 2010) setting, as is done in Anthony Lee (2011) which also contains an earlier version of the stopping time framework detailed in Section 6.2. Anthony Lee (2012) and Del Moral et al. (2015) provide some examples of each in simple scenarios.

Example 22 (Pseudo-marginal MTM)**.** In particular, in this setting one targets a distribution with density $\varpi(z)$ w.r.t. some measure ν where $\varpi : \mathsf{Z} \to \mathbb{R}_+$ cannot be calculated but for any $z \in \mathsf{Z}$ one can simulate a random variable $w \sim Q_z$ with expectation $\varpi(z)$. We introduce the auxiliary distribution with density

$$\pi(\mathrm{d}z, \mathrm{d}w) = \nu(\mathrm{d}z) w Q_z(\mathrm{d}w),$$

such that $\pi(\mathrm{d}z) = \nu(\mathrm{d}z) \int w Q_z(\mathrm{d}w) = \nu(\mathrm{d}z)\varpi(z)$. Letting Q be a Markov kernel evolving on Z, $\mathtt{W} = (w_1, \ldots, w_n) \in \mathbb{R}^n$ and $\mathtt{W}' = (w'_1, \ldots, w'_n) \in \mathbb{R}^n$ we consider the choice $\xi = (z, z', \mathtt{W}, \mathtt{W}', k, \ell) \in \mathsf{Z} \times \mathsf{Z} \times \mathbb{R}^n \times \mathbb{R}^n$, $\xi_0 = (z, w_k)$ and

$$\mu(\mathrm{d}z, \mathrm{d}z', \mathrm{d}\mathtt{W}, \mathrm{d}\mathtt{W}', k, \ell) = \frac{\mathbb{I}\{k \in \llbracket n \rrbracket\}}{n} \nu(\mathrm{d}z) w_k Q_z^{\otimes n}(\mathrm{d}\mathtt{W}) Q(z, \mathrm{d}z') Q_{z'}^{\otimes n}(\mathrm{d}\mathtt{W}') \varsigma(\ell; \mathtt{W}'),$$

where $\varsigma(k; \mathtt{W}) \propto w_k \mathbf{1}_{\llbracket n \rrbracket}(k)$. The involution can be chosen to be $\phi(z, z', \mathtt{W}, \mathtt{W}', k, \ell) = (z', z, \mathtt{W}', \mathtt{W}, \ell, k)$, giving the acceptance ratio

$$r(\xi) = \frac{p(z')q(z', z)}{p(z)q(z, z')} \cdot \frac{\sum_{i=1}^n w'_i}{\sum_{i=1}^n w_i},$$

where for simplicity we assume that $\{Q(z, \cdot); z \in \mathsf{Z}\}$ and ν have densities $\{q(z, \cdot); z \in \mathsf{Z}\}$ and p w.r.t. some dominating reference measure. We can view the averages of the w_i (resp. w'_i) as approximations of $\varpi(z)$ (resp. $\varpi(z')$) and the value of n controls the variability of the approximation. The corresponding Markov kernel is subtly different from the standard pseudo-marginal approach, in that here one simulates $k \sim \text{Uniform}(\llbracket n \rrbracket)$ and $\mathtt{W}_{-k} \sim Q_z^{\otimes n-1}$ rather than having these variables fixed. In some sense, one can view the standard pseudo-marginal kernel as a MwG approach where one fixes $(k, \mathtt{W})$, rather than only fixing ξ_0.

The next example is an interesting variant in which a shared stopping time is defined, and which has been shown to inherit desirable properties from the limiting MH kernel associated with the pair (π, Q) as $n \to \infty$ but which would naturally require computation of f under conditions where the kernel of Example 22 with any fixed n would not (Anthony Lee and Łatuszyński, 2014).

Example 23 (One-hit kernel of A. Lee, C. Andrieu, and A. Doucet, 2012)**.** Consider the setting of Example 22 but where Q_z is a Bernoulli($\varpi(z)$) distribution with $\varpi : \mathsf{Z} \to [0, 1]$. In this case, $w = 1$ under π. We want here to adapt the number of simulations so that the acceptance ratio is a reasonable approximation of the limiting acceptance ratio as $n \to \infty$, but does not require an excessive number of simulations. In particular, using a fixed number of simulations may lead to acceptance ratios with a large variance and hence a Markov chain that can get "stuck" for long periods when in regions of the state space with very small $\varpi(z)$. Let $\xi = (z, z', \mathtt{W}, \mathtt{W}', n) \in \mathsf{Z} \times \mathsf{Z} \times \{0, 1\}^{\mathbb{N}_*} \times \{0, 1\}^{\mathbb{N}_*} \times \mathbb{N}_*$, $\xi_0 = (z, w_1)$ and let for $z \in \mathsf{Z}$, F_z be the probability measure associated with an infinite sequence of independent Q_z-distributed random variables. The idea is given $z, z' \in \mathsf{Z}$ and $w_1 = 1$ we wish to simulate $w_i \overset{\text{iid}}{\sim} Q_z$, $i \in \mathbb{N}, i \geq 2$ and $w'_i \overset{\text{iid}}{\sim} Q_{z'}$, $i \in \mathbb{N}_*$ independently until there is one further "hit", i.e. w_i and/or w'_i is equal to 1. So we define

$$\tau(\xi) := \inf\{n \geq 1 : s_n(\mathtt{W}, \mathtt{W}') = 1\},$$

where $s_n(\mathtt{W}, \mathtt{W}') = \mathbb{I}\{\sum_{i=1}^n w_i + \sum_{i=1}^n w'_i \geq 2\}$. We then define

$$\mu(\mathrm{d}z, \mathrm{d}z', \mathrm{d}\mathtt{W}, \mathrm{d}\mathtt{W}', n) := \nu(\mathrm{d}z) w_1 Q(z, \mathrm{d}z') F_z(\mathrm{d}\mathtt{W}) F_{z'}(\mathrm{d}\mathtt{W}') \mathfrak{s}(n, \mathtt{W}, \mathtt{W}'),$$

where for $n \in \mathbb{N}_*$, $\mathfrak{s}(n, \mathtt{W}, \mathtt{W}') = s_n(\mathtt{W}, \mathtt{W}') \prod_{i=1}^{n-1} [1 - s_i(\mathtt{W}, \mathtt{W}')]$ so that if $\xi = (z, z', \mathtt{W}, \mathtt{W}', n) \sim \mu$ then $n = \tau(\xi)$. We define the involution

$$\phi(z, z', \mathtt{W}, \mathtt{W}', n) = (z', z, \sigma_n(\mathtt{W}'), \sigma_n(\mathtt{W}), n),$$

where for $i \in \mathbb{N}_*$, $\sigma_i \colon \{0, 1\}^{\mathbb{N}_*} \to \{0, 1\}^{\mathbb{N}_*}$ is the permutation that swaps its $1-$st and $i-$th inputs. We can obtain, for ξ such that $p(z)q(z, z') > 0$,

$$r(\xi) = \frac{p(z')q(z', z)}{p(z)q(z, z')} \mathbb{I}\{w_1 = w'_n = 1, \tau(\xi) = \tau \circ \phi(\xi) = n\},$$

i.e. it is essential that $w'_n = 1$ and that the stopping time is preserved by the involution. We observe that if $\tau(\xi) = 1$, then necessarily $w'_1 = 1$ and $\tau \circ \phi(\xi) = 1 = \tau(\xi)$, so the indicator above is 1. Now consider $\tau(\xi) > 1$. If $w_{\tau(\xi)} = 0$ then $(w_1, w_2, \ldots, w_{\tau(\xi)}) = (1, 0, \ldots, 0)$, while necessarily $(w'_1, \ldots, w'_{\tau(\xi)}) = (0, \ldots, 0, 1)$ so $\tau \circ \phi(\xi) = \tau(\xi)$ and the indicator above is 1. If on the other hand $w_{\tau(\xi)} = 1$ then $(w_1, w_2, \ldots, w_{\tau(\xi)}) = (1, 0, \ldots, 1)$ so even if $w'_{\tau(\xi)} = 1$ we have $\tau \circ \phi(\xi) = 1 \neq \tau(\xi)$ and so $r(\xi) = 0$.

Algorithm 9 Stochastic delayed rejection

(a) Given $z \in \mathsf{Z}$, set $k \leftarrow 0$ and $z_0 \leftarrow z$.

(b) Set $k \leftarrow k+1$ and simulate $z_k \sim Q_k(\mathsf{Z}^{k-1}, \cdot)$.

(c) With probability $\alpha_k(\mathsf{Z}^k)$ output $\phi_k(\mathsf{Z}^k)_0$, otherwise go to 2.

7 Delayed rejection

In delayed rejection, several sources of randomness and involutions are considered in turn until one is accepted.

7.1 Stochastic delayed rejection

Let π be a probability distribution on $(\mathsf{Z}_0, \mathscr{Z}_0)$. For $k \in \mathbb{N}_*$ let $(\mathsf{Z}_k, \mathscr{Z}_k)$ be measurable spaces, for $\mathsf{z}^{k-1} \in \mathsf{Z}^{k-1} := \mathsf{Z}_0 \times \mathsf{Z}_1 \times \cdots \mathsf{Z}_{k-1}$ let $Q_k(\mathsf{z}^{k-1}, \cdot)$ be probability distributions on $(\mathsf{Z}_k, \mathscr{Z}_k)$. Define $\eta_k(\mathrm{d}\mathsf{z}^k) := \pi(\mathrm{d}z_0) \prod_{i=1}^k Q_i(\mathsf{z}^{i-1}; \mathrm{d}z_i)$, for $k \in \mathbb{N}_*$ let $\phi_k \colon \mathsf{Z}^{k+1} \to \mathsf{Z}^{k+1}$ be involutions and let $\alpha_k = a_k \circ r_k$ for an acceptance function a_k and $\beta_k(\mathsf{z}^k) = \beta_{k-1}(\mathsf{z}^{k-1})[1 - \alpha_{k-1}(\mathsf{z}^{k-1})]$ with $\beta_0 \equiv 1$ and $\alpha_0 \equiv 0$ where

$$r_k(\mathsf{z}^k) := \begin{cases} \frac{\beta_k \circ \phi_k}{\beta_k}(\mathsf{z}^k) \frac{\mathrm{d}\eta_{k,S_k}^{\phi_k}}{\mathrm{d}\eta_{k,S_k}}(\mathsf{z}^k) & \mathsf{z}^k \in S_k \\ 0 & \text{otherwise} \end{cases},$$

with $S_k := S(\eta_k, \eta_k^{\phi_k}) \cap \{\mathsf{z}^k \in \mathsf{Z}^k \colon \beta_k(\mathsf{z}^k) \wedge \beta_k \circ \phi_k(\mathsf{z}^k) > 0\}$, where $S(\eta_k, \eta_k^{\phi_k})$ is as in Theorem 3. The delayed rejection algorithm is described in Alg. 9 and its justification follows from the following:

Proposition 5. *With the notation above, define the probability distribution of marginal π,*

$$\mu(k, \mathrm{d}\mathsf{z}^k) := \alpha_k(\mathsf{z}^k)\beta_k(\mathsf{z}^k)\eta_k(\mathrm{d}\mathsf{z}^k),$$

on $E = \{(k, \mathsf{z}^k) \in \{k\} \times \mathsf{Z}^k : k \in \mathbb{N}\}$ and for any $\xi = (k, \mathsf{z}^k) \in E$ the involution $\phi(\xi) = \phi(k, \mathsf{z}^k) = \big(k, \phi_k(\mathsf{z}^k)\big)$. Then,

$$r(\xi) = \begin{cases} 1 & \xi \in S(\mu, \mu^\phi) \\ 0 & \textit{otherwise} \end{cases}.$$

Although $\beta_k(\mathsf{z}^k)$ can be updated as the algorithm progresses, the computation of $\beta_k \circ \phi_k(\mathsf{z}^k)$ can be expensive. Indeed, letting $\zeta^k := \phi_k(\mathsf{z}^k)$ for $k \in \mathbb{N}_*$, we have

$$\beta_k \circ \phi_k(\mathsf{z}^k) = \beta_k(\zeta^k) = \prod_{i=1}^{k-1} [1 - \alpha_i(\zeta^i)],$$

where for $i \in [\![k-1]\!]$ $\alpha_i(\zeta^i) = a_i \circ r_i(\zeta^i)$ which may need to be computed afresh for each value of k in general. We will see in Subsection 7.2 an interesting scenario where this is not the case.

Example 24 (Delayed-rejection of Tierney and Mira, 1999)**.** Assume Q_i has density q_i with respect to the Lebesgue or counting measure for each $i \in [\![n]\!]$ for some $n \in \mathbb{N}_*$ and $\phi_i(z_1, z_2, \ldots, z_{i-1}, z_i) = (z_i, z_{i-1}, \ldots, z_2, z_1)$ for $i \in [\![n-1]\!]$ "reverse time" and $\phi_n = \mathrm{Id}$ (which ensures finite computations).

Example 25 (Generalized delayed-rejection of Green and Mira, 2001)**.** In this scenario involutions other than those of Example 24 can be used. As an example, one can choose $Q_1(x, \mathrm{d}y)$ on $\mathsf{X} \times \mathscr{Y}$, $Q_2(x, y; \mathrm{d}z, \mathrm{d}w)$ on $\mathsf{X} \times \mathsf{Y} \times \mathscr{Z} \otimes \mathscr{W}$ and Q_3 arbitrary. Assume Q_i has density q_i for each $i \in \{1, 2\}$, we may choose $\phi_1(x, y) = (y, x)$, $\phi_2(x, y, z, w) = (z, w, x, y)$ and $\phi_3(x, y, z, w, \ldots) = (x, y, z, w, \ldots)$.

Algorithm 10 Deterministic delayed rejection

(a) Given $z \in \mathsf{Z}$, set $k \leftarrow 0$.

(b) Set $k \leftarrow k+1$.

(c) With probability $\alpha_k(z)$ output $\phi_k(z)$ otherwise go to 2.

7.2 Deterministic delayed rejection

Delayed rejection can be usefully applied to sample from π defined on $(\mathsf{Z}, \mathscr{Z})$ using purely deterministic proposals. It is possible to use the framework above, but it is more convenient notationally and conceptually to instead consider $E := \{(k,z)\colon k \in \mathbb{N}, z \in \mathsf{Z}\}$, the embedding distribution, for $k \in \mathbb{N}^*$,

$$\mu(k, \mathrm{d}z) = \alpha_k(z)\beta_k(z)\pi(\mathrm{d}z),$$

involutions $\phi_k\colon \mathsf{Z} \to \mathsf{Z}$ and as before, for each $k \in \mathbb{N}$, let $\alpha_k = a_k \circ r_k$, $\beta_k(z) = \beta_{k-1}(z)[1-\alpha_{k-1}(z)]$ with $\beta_0 \equiv 1$ and $\alpha_0 \equiv 0$, where

$$r_k(z) = \begin{cases} \frac{\mathrm{d}\pi_{S_k}^{\phi_k}}{\mathrm{d}\pi_{S_k}}(z)\frac{\beta_k \circ \phi_k}{\beta_k}(z) & z \in S_k \\ 0 & \text{otherwise} \end{cases},$$

with $S_k = S(\pi, \pi^{\phi_k}) \cap \{z \in \mathsf{Z}\colon \beta_k(z) \wedge \beta_k \circ \phi_k(z) > 0\}$. An algorithmic presentation of delayed rejection with deterministic proposals is given in Alg. 10. Its justification follows along the same lines as above, and as before one can choose $\phi_n = \mathrm{Id}$ for some $n \in \mathbb{N}_*$ to ensure that the stage n "proposal" is accepted.

As in the stochastic scenario, the computation of $\beta_k \circ \phi_k(z)$ can be expensive since this requires in particular the computation of $r_1 \circ \phi_k(z), \ldots, r_{k-1} \circ \phi_k(z)$. Assume for simplicity that π has a density ϖ with respect to some measure ν invariant under ϕ_i for $i \in [\![k-1]\!]$, then we remark that on S_k

$$r_i \circ \phi_k(z) = \frac{\varpi \circ \phi_i \circ \phi_k}{\varpi \circ \phi_k}(z)\frac{\beta_i \circ \phi_i \circ \phi_k}{\beta_i \circ \phi_k}(z).$$

Then, if for $i \leq k$ the identity $\varpi \circ \phi_i \circ \phi_k = \varpi \circ \phi_{k-i}$ holds, we see that no new evaluation of the probability density is required, which is to be contrasted with the general setup in Subsection 7.1. This identity holds when for $\psi\colon \mathsf{Z} \to \mathsf{Z}$, assumed invertible, is such that for an involution $\sigma\colon \mathsf{Z} \to \mathsf{Z}$, $\sigma \circ \psi \circ \sigma = \psi^{-1}$ one considers the involutions $\phi_i = \sigma \circ \psi^i$ and has the property $\varpi \circ \sigma = \varpi$, since

$$\phi_i \circ \phi_k = \sigma \circ \psi^i \circ \sigma \circ \psi^k = \psi^{-i} \circ \psi^k = \psi^{k-i} = \sigma \circ \phi_{k-i}.$$

This is the setup considered in Sohl-Dickstein, Mudigonda, and DeWeese (2014) and Campos and J. Sanz-Serna (2015) where an additional twist, detailed in the next subsection, is used.

Example 26 (DR deterministic)**.** Consider π defined on $\left(\mathsf{X}\times\mathsf{V}, \mathscr{X}\otimes\mathscr{V}\right)$ of density $\varpi(x,v) = \gamma(x)\kappa(v)$ and let $\phi_i := \sigma \circ \psi^i$ for $i \in [\![n-1]\!]$ and $\phi_n = \mathrm{Id}$ with $\psi(x,v) := (x+v, v)$ and $\sigma(x,v) = (x,-v)$. This can be useful when trying to traverse a region of low probability. As an example, let $\mathsf{X}\times\mathsf{V} = [\![3]\!] \times \{-1,1\}$ and $\varpi(1,1) = \varpi(1,-1) > 0$, $\varpi(3,1) = \varpi(3,-1) > 0$ but $\varpi(2,1) = \varpi(2,-1) = 0$. In this scenario it is a good idea to choose $n = 3$ rather than the standard $n = 2$ choice.

7.3 Sliced delayed rejection

The introduction of an auxiliary slice variable can mitigate the computational cost of the delayed rejection approach, and some recently proposed algorithms Sohl-Dickstein, Mudigonda, and DeWeese

(2014) and Campos and J. Sanz-Serna (2015) can be viewed as following this principle. In particular, we can define

$$\varpi(z,u) = \varpi(z)\mathbb{I}\{u \leq \varpi(z)\}/\varpi(z) = \mathbb{I}\{u \leq \varpi(z)\},$$

and use a slice sampler (Neal, 2003) (see Definition 6), that is a MwG alternating between updating u given z and vice-versa, that is sampling uniformly from the "slice" $\{z \in \mathsf{Z} : \varpi(z) \geq u\}$, for a fixed $u \in \mathbb{R}_+$. One may use any Markov kernel that leaves this distribution invariant and we naturally focus on MH type updates.

Example 27 (Extra chance slice)**.** For some fixed u, let $\varpi_u(z) \propto \mathbb{I}\{u \leq \varpi(z)\}$. Let $\phi_i(z) = \sigma \circ \psi_i(z)$ for $i \in [\![n-1]\!]$ and $\phi_n = \mathrm{Id}$. If $a_i(r) = a(r) = 1 \wedge r$, we find that (see Appendix C for a proof) for $z \in \mathsf{Z}$, for $k \in [\![n]\!]$ and the convention $\vee_{i=1}^0 = 0$,

$$r_k(z) = \mathbb{I}\{\vee_{i=1}^{k-1} \varpi \circ \phi_i(z) < u \leq \varpi(z) \wedge \varpi \circ \phi_k(z)\}$$

while $\beta_k(z) = \mathbb{I}\{\varpi(z) \wedge \vee_{i=1}^{k-1} \varpi \circ \phi_i(z) < u\}$. Hence, one accepts as soon as $\varpi \circ \phi_k(z) \geq u$ or one reaches the identity involution $\phi_n = \mathrm{Id}$. One notices that for $\varpi(z) > 0$ and $u \sim \mathrm{Uniform}(0, \varpi(z))$ then $u_0 := u/\varpi(z) \sim \mathrm{Uniform}(0,1)$ and one can rewrite

$$r_k(z) = \mathbb{I}\Big\{\vee_{i=1}^{k-1} \varpi \circ \phi_i(z)/\varpi(z) < u_0 \leq 1 \wedge \big(\varpi \circ \phi_k(z)/\varpi(z)\big)\Big\}.$$

The overall slice sampler therefore looks like a standard MH algorithm targetting ϖ, where given $u_0 \sim \mathrm{Uniform}(0,1)$ one scans the states $\phi_i(z)$ for $i \in [\![n]\!]$ until the right hand side inequality is satisfied or n is reached. When $n = 2$ we recover the standard MH algorithm targetting ϖ and with deterministic proposal, corresponding to a remark going as far back as Higdon (1998).

In the context of HMC samplers this can be a way of taking into account the oscillatory nature of the energy $i \mapsto H \circ \psi^i(x,v)$ under the leapfrog dynamics. More specifically we may have $\varpi \circ \phi_k(z) \geq u$ even though $\varpi \circ \phi_i(z) < u$ for $i \in [\![k-1]\!]$. Note that ψ may involve several steps of the numerical integrator (which preserves Lebesgue measure and is time-reversible).

Example 28. The "sequential-proposal Metropolis(–Hastings) algorithm" of Park and Atchadé (2020) shares the precise structure of Campos and J. Sanz-Serna (2015) albeit in the scenario where the states are proposed randomly, but this connection was not made by the authors.

7.4 Discrete time bouncy particle samplers

Let b be a volume preserving "bounce" involution, e.g. with $\mathrm{b}_v(x,v) := v - 2\langle v, n(x)\rangle n(x)$ for some function $n : \mathbb{R}^d \to \mathbb{R}^d$ such that for all $x \in \mathsf{X}$, $\|n(x)\| = 1$ we let $\mathrm{b}(x,v) := \big(x, \mathrm{b}_v(x,v)\big)$ for $(x,v) \in \mathsf{X}\times\mathsf{V}$. To fix ideas, for the two following examples the scenario where $\psi(x,v) = (x+v, v)$ corresponds to the algorithms of Sherlock and Thiery (2017) and Vanetti et al. (2017). Similar ideas are briefly alluded to in Neal (2003).

Example 29 (Bouncy I - Sherlock and Thiery (2017))**.** Let $\phi_1 = \sigma \circ \psi$ and $\phi_2 = \phi_1 \circ \mathrm{b} \circ \phi_1$. Note that ϕ_2 is an involution since ϕ_1 and b are involutions. We have the convenient property that $\phi_1 \circ \phi_2 = \mathrm{b} \circ \phi_1$ (since ϕ_1 is an involution), and $\mathrm{b} \circ \phi_1(z)$ will already have been computed to produce $\phi_2(z)$. We have $\mathrm{b} \circ \phi_1(x,v) = \big(x+v, \mathrm{b}_v(x+v,-v)\big)$ and $\phi_2(x,v) = \big(x+v+\mathrm{b}_v(x+v,-v), -\mathrm{b}_v(x+v,-v)\big)$ and therefore for $\xi \in S$ the acceptance ratio is

$$r_2(\xi) = \frac{\bar{\alpha}_1 \circ \phi_2(x,v)(\gamma \otimes \kappa) \circ \phi_2(x,v)}{\bar{\alpha}_1(x,v)\gamma \otimes \kappa(x,v)} = \frac{\gamma\big(x+v+\mathrm{b}_v(x+v,-v)\big)}{\gamma(x)}$$

since,

$$r_1 \circ \phi_2(x,v) = \frac{\gamma(x+v)\kappa \circ \mathrm{b}_v(x+v,-v)}{\gamma(x)\kappa(v)} = r_1(x,v).$$

Example 30 (Bouncy II - Vanetti et al. (2017))**.** Let $\phi_1 = \sigma \circ \psi$ and $\phi_2 = \mathrm{b}$, where b is an involution. Here more computations are required since $\phi_1 \circ \phi_2 = \phi_1 \circ \mathrm{b}$ and $\phi_1 \circ \mathrm{b}(z)$ is not typically computed as a by-product of computing $\phi_1(z)$ or $\phi_2(z)$. Here for $\xi = (x, v) \in S$

$$r_2(\xi) = \frac{\bar{\alpha}_1 \circ \phi_2(x,v)(\gamma \otimes \kappa) \circ \phi_2(x,v)}{\bar{\alpha}_1(x,v)\gamma \otimes \kappa(x,v)} = \frac{\bar{\alpha}_1 \circ \phi_2(x,v)}{\bar{\alpha}_1(x,v)}$$

where

$$r_1 \circ \phi_2(x,v) = \frac{\gamma\big(x + \mathrm{b}_v(x,v)\big)\kappa\big(-\mathrm{b}_v(x+v,-v)\big)}{\gamma(x)\kappa(v)} = \frac{\gamma\big(x + \mathrm{b}_v(x,v)\big)}{\gamma(x)}.$$

7.5 Discrete-time exact event chain algorithms

In a lineage of contributions Jaster (1999), Bernard, Krauth, and Wilson (2009), Michel, Kapfer, and Krauth (2014), Michel, Mayer, and Krauth (2015), and Michel (2016) efficient continuous time nonreversible Markov process Monte Carlo (MPMC) algorithms have been developed to sample from models arising in statistical physics. We show here that it is possible to develop discrete time and exact counterparts of those, that is algorithms of finite run time without any approximation but the machine numerical precision limit and are ensured to leave the desired distribution invariant. More specifically let $x = (x_1, \ldots, x_m)$ for some $m \in \mathbb{N}$ where $x_1, \ldots, x_m \in \mathsf{X}$. For example X might be a bounded subset of $\mathbb{R}^d$ for some d or as is a common in the physics literature a torus. This can be thought of as the positions of m particles, modelled as spheres. We also define a velocity variable $v \in \mathsf{V} \subset \mathbb{R}^d$. The target distribution has density with respect to some measure ν, which can be the product of the Lebesgue or Hausdorff or counting measure depending on the scenario considered,

$$\varpi(x,v) = \kappa(v)\gamma(x) \propto \kappa(v) \prod_{1 \le i < j \le m} \mathbb{I}\left\{\|x_i - x_j\| > \delta_{ij}\right\}, \tag{14}$$

where $\delta_{ij} \in \mathbb{R}_+$ for all $i, j \in [\![m]\!]$, $i < j$ and it is assumed that $\kappa(v) = \kappa(-v)$ for all $v \in \mathsf{V}$. No simplify notation we introduce the feasible set $F \subset \mathsf{X}$ such that $\mathbb{I}\{x \in F\} = \prod_{1 \le i < j \le m} \mathbb{I}\left\{\|x_i - x_j\| > \delta_{ij}\right\}$. In the absence of mean field we see the necessity to constrain X to be "bounded" for this to define a probability distribution. This clearly accommodates hard constraints on the distance between the particles. We now briefly describe the aforementioned MPMC in the hard sphere scenario given by (14). This MPMC is a so-called piecewise deterministic Markov process where a sphere, labelled $i \in [\![m]\!]$, evolves continuously along a straight line of direction the velocity v until a collision with another sphere occurs, say $j \in [\![m]\!] \setminus \{i\}$ or until an exponential clock of fixed parameter rings. When a collision occurs the velocity is transferred to sphere j, while when the clock rings a new velocity is drawn afresh from κ. For soft potentials implementation of the algorithm will typically require time discretisation in order to determine the time to a "soft" collision. For completeness we provide a description of the generator of the MPMC above for soft potentials in Appendix E. In Alg. 11 we introduce a novel exact discretization of aforementioned MPMC which circumvents the need for a time discretization approximation thanks to a MH kernel involving delayed rejection. Note that in practice this kernel is composed with $\mathfrak{S}$ such that for any $f \in \mathsf{X}^m \times \mathsf{V} \times [\![m]\!]^2$, $\mathfrak{S}f(x,v,i,j) = f \circ \sigma(x,v,i,j)$ where for any $(x,v,i,j) \in \mathsf{X}^m \times \mathsf{V} \times [\![m]\!]^2$, $\sigma(x,v,i,j) = (x,-v,i,j)$ denotes the function that flips the sign of v.

To justify the algorithm we let $\xi := (x,v,i,j) \in \mathsf{X}^m \times \mathsf{V} \times [\![m]\!]^2$: here i is the index of the particle that is "moving" and j the index of a candidate particle that will be "given" the velocity of the ith particle, in a way that will become clear. The algorithm is a two-stage delayed rejection MH kernel. The first involution is $\phi_1 = \sigma \circ \psi_1$ where $\psi_1(x,v,i,j) = (x',v,i,j)$ with $x'_i = x_i + v$ and $x'_j = x_j$ for $j \in [\![m]\!] \setminus \{i\}$, which we may denote $x' = x + \mathbf{e}_i \circledast v$ with $\circledast$ the Kronecker product and $\{\mathbf{e}_i \in \mathsf{X}^m, i \in [\![m]\!]\}$ such that $(\mathbf{e}_i)_j = \mathbb{I}\{i = j\}$. An interpretation of ψ_1 is that the $i-$th particle is translated by v and all other particles remain fixed. Let $I(x,v,i) := \{j \in [\![m]\!] \setminus \{i\} : \|x_i + v - x_j\| \le \delta_{ij}\}$, i.e. $I(x,v,i)$ is the set of

Algorithm 11 $\mathfrak{S}-$ symmetrisation of the discrete time event chain kernel

Input: (x, v, i)

(a) Set $I \leftarrow \emptyset$. For $k \in [\![m]\!] \setminus \{i\}$, if $\|x_i + v - x_k\| \leq \delta_{ik}$ set $I \leftarrow I \cup \{k\}$.

(b) If $I = \emptyset$, set $j = i$, $x_i \leftarrow x_i + v$ and output $(x, -v, i, j)$.

(c) Otherwise, sample $j \sim \text{Uniform}(I)$.

(d) Set $I' \leftarrow \emptyset$. For $k \in [\![m]\!] \setminus \{j\}$, if $\|x_k + v - x_j\| \leq \delta_{jk}$ set $I' \leftarrow I' \cup \{k\}$.

(e) With probability $a(|I|/|I'|)$ output $(x, -v, j)$, otherwise output (x, v, i).

particle indices j such that $x_i + v$ "collides" with x_j and let $|I|(x, v, i) := |I(x, v, i)|$, so that $|I|(x, v, i)$ is the number of such collisions. The second involution is simply $\phi_2(x, v, i, j) = (x, -v, j, i)$, that is particle j becomes active and has velocity $-v$. We define $\mu(\mathrm{d}\xi) \propto \gamma \otimes \kappa(\mathrm{d}x, \mathrm{d}v)\mathbb{I}\{i \in [\![m]\!]\}q(j; x, v, i)$, where we are free to choose the following proposal distribution for the next active particle, among those in $I(x, v, i)$,

$$q(j; x, v, i) = \begin{cases} \mathbb{I}\{i = j\} & I(x, v, i) = \emptyset, \\ \frac{\mathbb{I}\{j \in I(x,v,i)\}}{|I|(x,v,i)} & I(x, v, i) \neq \emptyset. \end{cases}$$

We have (with $0/0 = 0$ here)

$$\rho(x, v, i, j) = \frac{\kappa(v)}{m}\mathbb{I}\{x \in F, i \in [\![m]\!]\} \begin{cases} \mathbb{I}\{I(x, v, i) = \emptyset, i = j\} & I(x, v, i) = \emptyset, \\ \frac{\mathbb{I}\{j \in I(x,v,i) \neq \emptyset\}}{|I|(x,v,i)} & I(x, v, i) \neq \emptyset. \end{cases}$$

Notice that $\mathbb{I}\{x + \mathbf{e}_i \circledast v \in F\} = \mathbb{I}\{I(x, v, i) = \emptyset\}$ and equivalently $\mathbb{I}\{x \in F\} = \mathbb{I}\{I \circ \tilde{\phi}_1(x, v, i) = \emptyset\}$ with $\tilde{\phi}_1(x, v, i) = (x + \mathbf{e}_i \circledast v, -v, i)$ and $I \circ \tilde{\phi}_1(x, v, i) := I\big(\tilde{\phi}_1(x, v, i)\big)$

$$\rho \circ \phi_1(x, v, i, j) = \frac{\kappa(-v)}{m}\mathbb{I}\{x + \mathbf{e}_i \circledast v \in F, i \in [\![m]\!]\} \begin{cases} \mathbb{I}\{x \in F, i = j\} & I(x, v, i) = \emptyset, \\ \frac{\mathbb{I}\{x \notin F, j \in I \circ \tilde{\phi}_1(x,v,i)\}}{|I| \circ \tilde{\phi}_1(x,v,i)} & I(x, v, i) \neq \emptyset. \end{cases}$$

where. Therefore, using that $\rho(x, v, i, j), \rho \circ \phi_1(x, v, i, j) \in \{0, \kappa(v)/m\}$ we obtain

$$\begin{aligned} r_1(x, v, i, j) &= \mathbb{I}\{x \in F, (x + \mathbf{e}_i \circledast v) \in F, i = j\} \\ &= \mathbb{I}\{I \circ \tilde{\phi}_1(x, v, i) = I(x, v, i) = \emptyset, i = j\}. \end{aligned}$$

For the second stage observe that for $i, j \in [\![m]\!]$ we have $\|x_i + v - x_j\| = \|x_i - (x_j - v)\|$ and therefore $\mathbb{I}\{j \in I(x, v, i) \neq \emptyset\} = \mathbb{I}\{i \in I(x, -v, j) \neq \emptyset\}$ and therefore

$$\rho \circ \phi_1(x, v, i, j) = 0 \iff \rho \circ \phi_1 \circ \phi_2(x, v, i, j) = 0$$

in which case

$$\rho \circ \phi_2(x, v, i, j) = \mathbb{I}\{x \in F\}\frac{\mathbb{I}\{i \in I(x, -v, j) \neq \emptyset\}}{|I|(x, -v, j)}\frac{\kappa(v)}{m}.$$

Therefore we conclude that

$$r_2(x, v, i, j) = \mathbb{I}\{x \in F, i \in I(x, -v, j) \neq \emptyset, j \in I(x, v, i) \neq \emptyset\}\frac{|I|(x, v, i)}{|I|(x, -v, j)},$$

and easy counterexamples show that there is no reason for the equality $|I|(x, v, i) = |I|(x, -v, j) \neq 0$ to hold in general, when the indicator function is one. In practice, one can implement the combination

of this kernel with the refreshments for (x, v, i) in various ways to save time. In particular, one may be able to determine the first time at which either a refreshment occurs or there is a collision. We do not consider these details here.

We now show that Alg. 11 can be straightforwardly adapted to accommodate "soft potentials" (or constraints) by using a slice sampler strategy and hence the introduction of instrumental variables. For example, assume that

$$\gamma(x) = \prod_{1 \le i < j \le m} \gamma_{ij}(x),$$

where $\gamma_{ij}(x) = \Gamma(\|x_i - x_j\|)$ with $\Gamma \colon \mathbb{R}_+ \to \mathbb{R}_+$ is non-decreasing and such that γ is a probability density on X for the relevant dominating measure. Then one can consider the instrumental distribution, with $u = (u_{ij}) \in \mathbb{R}_+^{m(m-1)/2}$ for $i \in [\![m-1]\!]$ and $j \in [\![i+1, m]\!]$,

$$\begin{aligned}
\varpi(x, v, u) &:= \gamma(x) \prod_{1 \le i < j \le m} \frac{\mathbb{I}\{u_{ij} \le \gamma_{ij}(x)\}}{\gamma_{ij}(x)} \\
&= \prod_{1 \le i < j \le m} \mathbb{I}\{u_{ij} \le \gamma_{ij}(x)\} \\
&= \prod_{1 \le i < j \le m} \mathbb{I}\left\{\|x_i - x_j\| \ge \Gamma^{-1}(u_{ij})\right\},
\end{aligned}$$

where $\Gamma^{-1}(u) := \inf\{y : \Gamma(y) \ge u\}$. Hence, for a fixed u, we have $\pi_u(x, v)$ of the same form as (14) with $\delta_{ij} = \Gamma^{-1}(u_{ij})$, suggesting the use of a MwG strategy to sample from π. It is naturally possible to consider more general forms for the γ_{ij} and adaptation of the algorithm is straightforward.

8 Acknowledgements

CA and SL acknowledge support from EPSRC "Intractable Likelihood: New Challenges from Modern Applications (ILike)" (EP/K014463/1). CA and AL acknowledge support of EPSRC grant CoSInES (EP/R034710/1) and CA acknowledges support of EPSRC grant Bayes4Health (EP/R018561/1).

References

Andrieu, Christophe (2016). "On random-and systematic-scan samplers". In: *Biometrika* 103.3, pp. 719–726.

– (Aug. 2019). "Slides of Lecture series at HSE conference Structural Inference in High-Dimensional Models 2, St. Petersburg, 26-30 August 2019". In: https://www.dropbox.com/sh/otc6oadsrxo1ggu/AADzkheTcavMx0A to slides. eprint: https://cs.hse.ru/hdilab/sihdm/2019/.

Andrieu, Christophe, Arnaud Doucet, and Roman Holenstein (2010). "Particle markov chain monte carlo methods". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.3, pp. 269–342.

Andrieu, Christophe, Arnaud Doucet, Sinan Yıldırım, et al. (2020). "Metropolis-Hastings with Averaged Acceptance Ratios". In: *ArXiv e-prints*.

Andrieu, Christophe and Samuel Livingstone (June 14, 2019). "Peskun-Tierney ordering for Markov chain and process Monte Carlo: beyond the reversible scenario". In: arXiv: http://arxiv.org/abs/1906.06197v1 [math

Andrieu, Christophe and Gareth O Roberts (2009). "The pseudo-marginal approach for efficient Monte Carlo computations". In: *The Annals of Statistics*, pp. 697–725.

Andrieu, Christophe and Johannes Thoms (2008). "A tutorial on adaptive MCMC". In: *Statistics and computing* 18.4, pp. 343–373.

Beaumont, Mark A. (2003). "Estimation of Population Growth or Decline in Genetically Monitored Populations". In: *Genetics* 164.3, pp. 1139–1160. ISSN: 0016-6731. eprint: https://www.genetics.org/content/164/3/1 URL: https://www.genetics.org/content/164/3/1139.

Bernard, Etienne P, Werner Krauth, and David B Wilson (2009). "Event-chain Monte Carlo algorithms for hard-sphere systems". In: *Physical Review E* 80.5, p. 056704.

Besag, Julian (1994). "Discussion of paper by Ulf Grenander and Michael I Miller". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 56.4, pp. 549–603.

Besag, Julian et al. (1995). "Bayesian computation and stochastic systems". In: *Statistical science*, pp. 3–41.

Betancourt, Michael (2017). "A conceptual introduction to Hamiltonian Monte Carlo". In: *arXiv preprint arXiv:1701.02434*.

Billingsley, Patrick (1995). *Probability and measure*. John Wiley & Sons.

Campos, Cédric M and JM Sanz-Serna (2015). "Extra chance generalized hybrid Monte Carlo". In: *Journal of Computational Physics* 281, pp. 365–374.

Ceperley, DM and M Dewing (1999). "The penalty method for random walks with uncertain energies". In: *The Journal of chemical physics* 110.20, pp. 9812–9820.

Chen, Ting-Li and Chii-Ruey Hwang (2013). "Accelerating reversible Markov chains". In: *Statistics & Probability Letters* 83.9, pp. 1956–1962.

Cusumano-Towner, Marco, Alexander K. Lew, and Vikash K. Mansinghka (2020). *Automating Involutive MCMC using Probabilistic and Differentiable Programming*. arXiv: 2007.09871 [stat.CO].

Del Moral, Pierre et al. (2015). "The alive particle filter and its use in particle Markov chain Monte Carlo". In: *Stochastic Analysis and Applications* 33.6, pp. 943–974.

Diaconis, Persi, Susan Holmes, and Radford M Neal (2000). "Analysis of a nonreversible Markov chain sampler". In: *Annals of Applied Probability*, pp. 726–752.

Durmus, Alain, Arnaud Guillin, and Pierre Monmarché (July 2018). "Piecewise Deterministic Markov Processes and their invariant measure". In: *arXiv e-prints*, arXiv:1807.05421, arXiv:1807.05421. arXiv: 1807.05421 [math.PR].

Durmus, Alain, Eric Moulines, and Eero Saksman (2017). "On the convergence of hamiltonian monte carlo". In: *arXiv preprint arXiv:1705.00166*.

Durrett, Rick (2019). *Probability: Theory and Examples*. 5th ed. Cambridge University Press.

Dutta, Somak and Sourabh Bhattacharya (2014). "Markov chain Monte Carlo based on deterministic transformations". In: *Statistical Methodology* 16, pp. 100–116.

Engelbert, H. and A. Shiryaev (1980). "On absolute continuity and singularity of probability measures". eng. In: *Banach Center Publications* 6.1, pp. 121–132. URL: http://eudml.org/doc/209108.

Fang, Youhan, Jesus-Maria Sanz-Serna, and Robert D Skeel (2014). "Compressible generalized hybrid Monte Carlo". In: *The Journal of chemical physics* 140.17, p. 174108.
Folland, Gerald B (1999). *Real analysis: modern techniques and their applications*. Vol. 40. John Wiley & Sons.
Fremlin, DH (Jan. 2010). *Measure Theory*. Second Edition. Vol. 2. https://wiki.math.ntnu.no/_media/tma4225/2011/f Torres Fremlin.
Glatt-Holtz, Nathan E., Justin A. Krometis, and Cecilia F. Mondaini (2020). *On the accept-reject mechanism for Metropolis-Hastings algorithms*. arXiv: 2011.04493 [math.ST].
Graham, Matthew McKenzie (2018). "Auxiliary Variable Markov Chain Monte Carlo Methods". In: *https://matt-graham.github.io/files/phd_thesis.pdf*.
Green, Peter J (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination". In: *Biometrika* 82.4, pp. 711–732.
Green, Peter J and Antonietta Mira (2001). "Delayed rejection in reversible jump Metropolis–Hastings". In: *Biometrika* 88.4, pp. 1035–1053.
Gustafson, Paul (1998). "A guided walk Metropolis algorithm". In: *Statistics and computing* 8.4, pp. 357–364.
Hairer, Martin, Andrew M Stuart, Sebastian J Vollmer, et al. (2014). "Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions". In: *The Annals of Applied Probability* 24.6, pp. 2455–2490.
Harland, Julian et al. (2017). "Event-chain Monte Carlo algorithms for three-and many-particle interactions". In: *EPL (Europhysics Letters)* 117.3, p. 30001.
Hastings, W Keith (1970). "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1, pp. 97–109.
Higdon, David M. (1998). "Auxiliary variable methods for Markov chain Monte Carlo with applications". In: *Journal of the American Statistical Association* 93.442, pp. 585–595. DOI: 10.1080/01621459.1998.10473712. eprint: https://amstat.tandfonline.com/doi/pdf/10.1080/01621459.1998.10473712. URL: https://amstat.tand
Hoffman, Matthew D and Andrew Gelman (2014). "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." In: *J. Mach. Learn. Res.* 15.1, pp. 1593–1623.
Horowitz, Alan M (1991). "A generalized guided Monte Carlo algorithm". In: *Physics Letters B* 268.2, pp. 247–252.
Jaster, Andreas (1999). "An improved Metropolis algorithm for hard core systems". In: *Physica A: Statistical Mechanics and its Applications* 264.1, pp. 134–141.
Lamb, Jeroen SW and John AG Roberts (1998). "Time-reversal symmetry in dynamical systems: a survey". In: *Physica D: Nonlinear Phenomena* 112.1, pp. 1–39.
Lee, A., C. Andrieu, and A. Doucet (2012). "Discussion of paper by P. Fearnhead and D. Prangle". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 74.3, pp. 419–474.
Lee, Anthony (2011). "On auxiliary variables and many-core architectures in computational statistics". PhD thesis. University of Oxford.
– (2012). "On the choice of MCMC kernels for approximate Bayesian computation with SMC samplers". In: *Proceedings of the 2012 Winter Simulation Conference (WSC)*. IEEE, pp. 1–12.
Lee, Anthony and Krzysztof Łatuszyński (2014). "Variance bounding and geometric ergodicity of Markov chain Monte Carlo kernels for approximate Bayesian computation". In: *Biometrika* 101.3, pp. 655–671.
Liu, Jun S, Faming Liang, and Wing Hung Wong (2000). "The multiple-try method and local optimization in Metropolis sampling". In: *Journal of the American Statistical Association* 95.449, pp. 121–134.
Maire, Florian, Randal Douc, and Jimmy Olsson (2014). "Comparison of asymptotic variances of inhomogeneous Markov chains with application to Markov chain Monte Carlo methods". In: *The Annals of Statistics* 42.4, pp. 1483–1510.
Metropolis, Nicholas et al. (1953). "Equation of state calculations by fast computing machines". In: *The journal of chemical physics* 21.6, pp. 1087–1092.

Michel, Manon (2016). "Irreversible Markov chains by the factorized Metropolis filter : algorithms and applications in particle systems and spin models". 2016PSLEE039. PhD thesis. URL: http://www.theses.fr/2016PSLEE
Michel, Manon, Sebastian C Kapfer, and Werner Krauth (2014). "Generalized event-chain Monte Carlo: Constructing rejection-free global-balance algorithms from infinitesimal steps". In: *The Journal of chemical physics* 140.5, p. 054116.
Michel, Manon, Johannes Mayer, and Werner Krauth (2015). "Event-chain Monte Carlo for classical continuous spin models". In: *EPL (Europhysics Letters)* 112.2, p. 20003.
Neal, Radford M (1994). "An improved acceptance procedure for the hybrid Monte Carlo algorithm". In: *Journal of Computational Physics* 111.1, pp. 194–203.
– (1996). "Sampling from multimodal distributions using tempered transitions". In: *Statistics and computing* 6.4, pp. 353–366.
– (1998). "Suppressing random walks in Markov chain Monte Carlo using ordered overrelaxation". In: *Learning in graphical models*. Springer, pp. 205–228.
– (2003). "Slice sampling". In: *Annals of statistics*, pp. 705–741.
– (2004). "Improving asymptotic variance of MCMC estimators: Non-reversible chains are better". In: *arXiv preprint math/0407281*.
– (2005). "Taking bigger Metropolis steps by dragging fast variables". In: *arXiv preprint math/0502099*.
Neal, Radford M et al. (2011). "MCMC using Hamiltonian dynamics". In: *Handbook of Markov Chain Monte Carlo* 2, pp. 113–162.
Neklyudov, Kirill et al. (2020). "Involutive mcmc: a unifying framework". In: *International Conference on Machine Learning*. PMLR, pp. 7273–7282.
Park, Joonha and Yves Atchadé (2020). "Markov chain Monte Carlo algorithms with sequential proposals". In: *Statistics and Computing* 30.5, pp. 1325–1345.
Peskun, P. H. (1973). "Optimum Monte-Carlo Sampling Using Markov Chains". In: *Biometrika* 60.3, pp. 607–612. ISSN: 00063444. URL: http://www.jstor.org/stable/2335011.
Poncet, Romain (2017). "Generalized and hybrid MCMC overdamped Langevin algorithms". In: *arXiv preprint arXiv:1701.05833*.
Sherlock, Chris and Alexandre H Thiery (2017). "A Discrete Bouncy Particle Sampler". In: *arXiv preprint arXiv:1707.05200*.
Sohl-Dickstein, Jascha, Mayur Mudigonda, and Michael R DeWeese (2014). "Hamiltonian Monte Carlo without detailed balance". In: *arXiv preprint arXiv:1409.5191*.
Sun, Yi, Jürgen Schmidhuber, and Faustino Gomez (2010). "Improving the asymptotic performance of Markov chain Monte-Carlo by inserting vortices". In: *Advances in Neural Information Processing Systems* 23, pp. 2235–2243.
Thin, Achille, Alain Durmus, et al. (2020). "Nonreversible MCMC from conditional invertible transforms: a complete recipe with convergence guarantees".
Thin, Achille, Nikita Kotelevskii, et al. (2020). *MetFlow: A New Efficient Method for Bridging the Gap between Markov Chain Monte Carlo and Variational Inference*. arXiv: 2002.12253 [stat.ML].
Tierney, Luke (Feb. 1998). "A note on Metropolis-Hastings kernels for general state spaces". In: *Ann. Appl. Probab.* 8.1, pp. 1–9. DOI: 10.1214/aoap/1027961031. URL: http://dx.doi.org/10.1214/aoap/1027961031.
Tierney, Luke and Antonietta Mira (1999). "Some adaptive Monte Carlo methods for Bayesian inference". In: *Statistics in medicine* 18.1718, pp. 2507–2515.
Vanetti, Paul et al. (2017). "Piecewise Deterministic Markov Chain Monte Carlo". In: *arXiv preprint arXiv:1707.05296*.
Yaglom, Akiva Moiseevich (1949). "On the statistical reversibility of Brownian motion". In: *Matematicheskii Sbornik* 66.3, pp. 457–492.

A Proofs

Proof of Theorem 3. Let $\lambda = \mu + \mu^\phi$ and define $\rho = \mathrm{d}\mu/\mathrm{d}\lambda$. We observe that $\lambda^\phi = \lambda$. Then for any $A \in \mathscr{E}$, we find using Theorems 2–1,

$$\begin{aligned}\int \mathbf{1}_A(\xi)\rho\circ\phi(\xi)\lambda(\mathrm{d}\xi) &= \int \mathbf{1}_A\circ\phi(\xi)\rho(\xi)\lambda^\phi(\mathrm{d}\xi)\\ &= \int \mathbf{1}_A\circ\phi(\xi)\mu(\mathrm{d}\xi)\\ &= \int \mathbf{1}_A(\xi)\mu^\phi(\mathrm{d}\xi),\end{aligned}$$

so $\rho\circ\phi = \mathrm{d}\mu^\phi/\mathrm{d}\lambda$. Define $S = \{\xi \in E : \rho(\xi) \wedge \rho\circ\phi(\xi) > 0\}$, which satisfies $\phi(S) = S$. Since S is the intersection of two measurable sets, it is measurable and so the restrictions of μ and μ^ϕ to S are well defined, and for $\xi \in S$,

$$\frac{\mathrm{d}\mu_S^\phi}{\mathrm{d}\mu_S}(\xi) = \frac{\rho\circ\phi}{\rho}(\xi), \qquad \frac{\mathrm{d}\mu_S}{\mathrm{d}\mu_S^\phi}(\xi) = \frac{\rho}{\rho\circ\phi}(\xi),$$

so $\mu_S \equiv \mu_S^\phi$. Let $A = \{\xi : \rho(\xi) = 0\}$ and $B = \{\xi : \rho(\xi) > 0, \rho\circ\phi(\xi) = 0\}$. We deduce that $\mu(A) = \int_A \rho(\xi)\lambda(\mathrm{d}\xi) = 0$ and $\mu^\phi(B) = \int_B \rho\circ\phi(\xi)\lambda(\mathrm{d}\xi) = 0$. Since $A \cap B = \emptyset$, $A \cup B = S^{\complement}$ and $\mu(A) = \mu^\phi(B) = 0$ we conclude that μ and μ^ϕ are mutually singular on $S^{\complement}$.

For part b(i), $(\rho\circ\phi/\rho)\circ\phi = \rho/\rho\circ\phi$ on S, since ϕ is an involution. Hence, $r\circ\phi = 1/r$ on S and since $a(0) = 0$, $\alpha = r\cdot\alpha\circ\phi$ on S and $\alpha = 0$ on $S^{\complement}$ by the definition of α and condition on a.

For part b(ii), combining the first part with $a(0) = 0$, Theorems 2–1 and ϕ an involution, we obtain

$$\begin{aligned}\int_E F(\xi)G\circ\phi(\xi)\alpha(\xi)\mu(\mathrm{d}\xi) &= \int_S F(\xi)G\circ\phi(\xi)\alpha(\xi)\mu_S(\mathrm{d}\xi)\\ &= \int_S F(\xi)G\circ\phi(\xi)r(\xi)\alpha\circ\phi(\xi)\mu_S(\mathrm{d}\xi)\\ &= \int_S F(\xi)G\circ\phi(\xi)\alpha\circ\phi(\xi)\mu_S^\phi(\mathrm{d}\xi)\\ &= \int_S F\circ\phi(\xi)G(\xi)\alpha(\xi)\mu_S(\mathrm{d}\xi)\\ &= \int_E F\circ\phi(\xi)G(\xi)\alpha(\xi)\mu(\mathrm{d}\xi).\end{aligned}$$

For the part b(iii), we define the sub-Markov kernels

$$T(\xi, A) = \alpha(\xi)\mathbf{1}_A(\phi(\xi)), \qquad \xi \in E, A \in \mathscr{E},$$

and

$$R(\xi, A) = \{1 - \alpha(\xi)\}\,\mathbf{1}_A(\xi), \qquad \xi \in E, A \in \mathscr{E},$$

so that $\Pi = T + R$. First we observe that

$$\int F(\xi)G(\xi')\mu(\mathrm{d}\xi)R(\xi,\mathrm{d}\xi') = \int F(\xi)G(\xi)\{1-\alpha(\xi)\}\mu(\mathrm{d}\xi) = \int G(\xi)F(\xi')\mu(\mathrm{d}\xi)R(\xi,\mathrm{d}\xi').$$

Then from the second part,

$$\begin{aligned}\int F(\xi)G(\xi')\mu(\mathrm{d}\xi)T(\xi,\mathrm{d}\xi') &= \int F(\xi)G\circ\phi(\xi)\alpha(\xi)\mu(\mathrm{d}\xi)\\ &= \int F\circ\phi(\xi)G(\xi)\alpha(\xi)\mu(\mathrm{d}\xi)\\ &= \int G(\xi)F(\xi')\mu(\mathrm{d}\xi)T(\xi,\mathrm{d}\xi').\end{aligned}$$

□

Proof of Proposition 1. Let $f, g : \mathsf{Z} \to [0,1]$ be measurable and let $F, G\colon \Xi \to [0,1]$ such that $F(\xi) = f(\xi_0)$ and $G(\xi) = g(\xi_0)$. Using reversibility of Π, we find

$$\begin{aligned}\int f(\xi_0)g(\xi_0')\pi(\mathrm{d}\xi_0)P(\xi_0,\mathrm{d}\xi_0') &= \int f(\xi_0)g(\xi_0')\pi(\mathrm{d}\xi_0)\mu_{\xi_0}(\mathrm{d}\xi_{-0})\Pi(\xi;\mathrm{d}\xi')\\ &= \int F(\xi)G(\xi')\mu(\mathrm{d}\xi)\Pi(\xi;\mathrm{d}\xi')\\ &= \int G(\xi)F(\xi')\mu(\mathrm{d}\xi)\Pi(\xi;\mathrm{d}\xi')\\ &= \int g(\xi_0)f(\xi_0')\pi(\mathrm{d}\xi_0)P(\xi_0,\mathrm{d}\xi_0').\end{aligned}$$

□

Proof of Lemma 1. Since g is an involution and $g(\mathsf{Y}) \subseteq \mathsf{Y}$, we have $\lambda_Y = \lambda_Y^g$ as explained in Remark 8. Let ψ be a non-negative function and $\lambda(\mathrm{d}x,\mathrm{d}y) = \lambda_X(\mathrm{d}x)\lambda_Y(\mathrm{d}y)$. Then we find

$$\begin{aligned}\int \psi(x,y)\lambda(\mathrm{d}x,\mathrm{d}y) &= \int \psi(x,y)\lambda_Y(\mathrm{d}y)\lambda_X(\mathrm{d}x)\\ &= \int \psi(x,y)\lambda_Y^g(\mathrm{d}y)\lambda_X(\mathrm{d}x)\\ &= \int \psi(x,g(y))\lambda_Y(\mathrm{d}y)\lambda_X(\mathrm{d}x)\\ &= \int \psi(f(x,y),g(y))\left|\det f_y'(x)\right|\lambda_Y(\mathrm{d}y)\lambda_X(\mathrm{d}x)\\ &= \int \psi\circ\phi(x,y)\left|\det f_y'(x)\right|\lambda(\mathrm{d}x,\mathrm{d}y)\end{aligned}$$

where $f_y(x) = x \mapsto f(x,y)$ for each $y \in \mathsf{Y}$. Since

$$\int \psi\circ\phi(x,y)\lambda^{\phi}(\mathrm{d}x,\mathrm{d}y) = \int \psi(x,y)\lambda(\mathrm{d}x,\mathrm{d}y),$$

and for an arbitrary, measurable, non-negative $g : E \to \mathbb{R}$ we can take $\psi = g\circ\phi^{-1}$ to obtain $g = \psi\circ\phi$, we obtain that $\mathrm{d}\lambda^{\phi}/\mathrm{d}\lambda = \left|\det f_y'(x)\right|$. □

Proof of Proposition 3. For part (a), the identity $\psi^{-1} = \sigma\circ\psi\circ\sigma$ is verified by observing that, since $\psi = \sigma\circ\phi$, $\sigma\circ\psi\circ\sigma = \phi\circ\sigma$ and that indeed $\phi\circ\sigma\circ\psi = \psi\circ\phi\circ\sigma = \mathrm{Id}$. We then note that $\phi = \sigma\circ\psi$ and so for $A \in \mathscr{E}$,

$$\lambda^{\phi}(A) = \lambda(\phi^{-1}(A)) = \lambda(\phi(A) = \lambda(\sigma\circ\psi(A)) = \lambda^{\sigma}(\psi(A)) = \lambda(\psi(A)) = \lambda^{\psi^{-1}}(A).$$

Since $\lambda^\sigma = \lambda$, for f integrable w.r.t. μ^σ,

$$\begin{aligned}\int f(\xi)\mu^\sigma(\mathrm{d}\xi) &= \int f\circ\sigma(\xi)\mu(\mathrm{d}\xi)\\ &= \int f\circ\sigma(\xi)\rho(\xi)\lambda(\mathrm{d}\xi)\\ &= \int f\circ\sigma(\xi)\rho(\xi)\lambda^\sigma(\mathrm{d}\xi)\\ &= \int f(\xi)\rho\circ\sigma(\xi)\lambda(\mathrm{d}\xi),\end{aligned}$$

and so $\mathrm{d}\mu^\sigma/\mathrm{d}\lambda = \rho\circ\sigma$. Since $\mu^\sigma = \mu$, we have $\rho\circ\sigma = \rho$. Hence, $\rho\circ\phi = \rho\circ\sigma\circ\psi = \rho\circ\psi$. We proceed to part (b), and note that

$$\Pi(\xi,\mathrm{d}\xi') = \alpha(\xi)\delta_{\phi(\xi)}(\mathrm{d}\xi') + [1-\alpha(\xi)]\delta_\xi(\mathrm{d}\xi'),$$

from which for $\xi\in E$ and $f\colon E\to[0,1]$, with $\Phi f := f\circ\phi$,

$$\Pi\mathfrak{S}f = \alpha(\xi)\cdot\Phi\mathfrak{S}f + [1-\alpha(\xi)]\mathfrak{S}^2 f$$

and use that $\mathfrak{S}^2 = \mathrm{Id}$ and $\Phi\mathfrak{S}f(\xi) = \phi\left(\mathfrak{S}f\right)(\xi) = \phi\left(f\circ\sigma\right)(\xi) = f\circ\sigma\circ\phi(\xi)$. Using the identities from part (a), we can use the general acceptance ratio for Π from Proposition 2 to obtain,

$$r(\xi) = \frac{\rho\circ\phi}{\rho}(\xi)\frac{\mathrm{d}\lambda^\phi}{\mathrm{d}\lambda}(\xi) = \frac{\rho\circ\psi}{\rho}(\xi)\frac{\mathrm{d}\lambda^{\psi^{-1}}}{\mathrm{d}\lambda}(\xi), \qquad \xi\in S.$$

For part (c), for $f,g\in E\to[0,1]$ we use that $\mathit{\Pi}\mathfrak{S} := \Pi$ satisfies detailed balance and $\mu\mathfrak{S} = \mu$

$$\begin{aligned}\int f(\xi)g(\xi')\mu(\mathrm{d}\xi)\mathit{\Pi}(\xi,\mathrm{d}\xi') &= \int f(\xi)\mathfrak{S}g(\xi')\mu(\mathrm{d}\xi)\mathit{\Pi}\mathfrak{S}(\xi,\mathrm{d}\xi')\\ &= \int f(\xi)\mathfrak{S}g(\xi')\mu(\mathrm{d}\xi')\mathit{\Pi}\mathfrak{S}(\xi',\mathrm{d}\xi)\\ &= \int f(\xi)\mathfrak{S}g(\xi')\mu\mathfrak{S}(\mathrm{d}\xi')\mathit{\Pi}\mathfrak{S}(\xi',\mathrm{d}\xi)\\ &= \int f(\xi)g(\xi')\mu(\mathrm{d}\xi')\mathfrak{S}\mathit{\Pi}\mathfrak{S}(\xi',\mathrm{d}\xi).\end{aligned}$$

For part (e) we proceed as above and for $\xi\in S$ notice that

$$\begin{aligned}\mathfrak{S}\left[\frac{\rho\circ\phi}{\rho}\frac{\mathrm{d}\lambda^\phi}{\mathrm{d}\lambda}\right](\xi) &= \frac{\rho\circ\phi\circ\sigma}{\rho\circ\sigma}(\xi)\frac{\mathrm{d}\lambda^\phi}{\mathrm{d}\lambda}\circ\sigma(\xi)\\ &= \frac{\rho\circ\phi\circ\sigma}{\rho}(\xi)\frac{\mathrm{d}\lambda^{\phi\circ\sigma}}{\mathrm{d}\lambda}(\xi),\end{aligned}$$

where we have used that

$$\begin{aligned}\int f(\xi)\frac{\mathrm{d}\lambda^{\phi}}{\mathrm{d}\lambda}\circ\sigma(\xi)\lambda(\mathrm{d}\xi) &= \int f(\xi)\frac{\mathrm{d}\lambda^{\phi}}{\mathrm{d}\lambda}\circ\sigma(\xi)\lambda^{\sigma}(\mathrm{d}\xi)\\ &= \int f(\xi)\frac{\mathrm{d}\lambda^{\phi}}{\mathrm{d}\lambda}(\xi)\lambda(\mathrm{d}\xi)\\ &= \int f\circ\phi(\xi)\lambda(\mathrm{d}\xi)\\ &= \int f\circ\phi(\xi)\lambda^{\sigma}(\mathrm{d}\xi)\\ &= \int f\circ\phi\circ\sigma(\xi)\lambda(\mathrm{d}\xi)\\ &= \int f(\xi)\frac{\mathrm{d}\lambda^{\phi\circ\sigma}}{\mathrm{d}\lambda}\lambda(\mathrm{d}\xi),\end{aligned}$$

where we have used that $\lambda^{\phi\circ\sigma}(A) = \lambda\big(\sigma\circ\phi(A)\big) = \lambda\big(\phi(A)\big) = \lambda^{\phi}\big(A\big)$. □

Proof of Lemma 2. By the definition of ψ, for measurable $A \in \mathsf{X}\times\mathsf{Y}$

$$\psi^{-1}(A) = \{(x,y) : (x, y + f(x)) \in A.$$

Let λ_X and λ_Y be, respectively, the Lebesgue measures on $\mathbb{R}^{d_X}$ and $\mathbb{R}^{d_Y}$. Using the translation-invariance of the Lebesgue measure, we obtain that for arbitrary, measurable A,

$$\begin{aligned}\lambda(\psi^{-1}(A)) &= \int_{\mathbb{R}^{d_X}}\int_{\{y:(x,y+f(x))\in A\}}\lambda_Y(\mathrm{d}y)\lambda_X(\mathrm{d}x)\\ &= \int_{\mathbb{R}^{d_X}}\int_{\{y:(x,y)\in A\}}\lambda_Y(\mathrm{d}y)\lambda_X(\mathrm{d}x)\\ &= \lambda(A),\end{aligned}$$

from which we can conclude that $\lambda^{\psi} = \lambda$. □

Proof of Lemma 3. Let λ denote the Lebesgue measure on $\mathbb{R}^{2d}$. By Lemma 2, ψ_A and ψ_B each preserve λ, and hence ψ preserves λ as a composition of λ-preserving maps. We observe that

$$\sigma\circ\psi(x,v) = \big(x + \jmath(v+\imath(x)), -v-\imath(x)-\imath[x+\jmath(v+\imath(x))]\big),$$

so that

$$\psi_B\circ\sigma\circ\psi(x,v) = \big(x+\jmath[v+\imath(x)], -v-\imath(x)\big),$$

and using $\jmath(-v') = -\jmath(v')$,

$$\psi_A\circ\psi_B\circ\sigma\circ\psi(x,v) = \big(x, -v-\imath(x)\big).$$

It follows that

$$\psi\circ\sigma\circ\psi(x,v) = (x,-v),$$

and so $\sigma\circ\psi\circ\sigma\circ\psi(x,v) = (x,v)$, from which we conclude that $\psi^{-1} = \sigma\circ\psi\circ\sigma$. □

Proof of Lemma 4. Let $n = \tau(\mathtt{Z})$, so $s_k(\mathtt{Z}) = 0$ for $k \in [\![n-1]\!]$ and $s_n(\mathtt{Z}) = 1$. From (10) it is sufficient to show that $s_k\circ\sigma_l(\mathtt{Z}) = s_k(\mathtt{Z})$ for

$$(k,l)\in[\![n]\!]\times[\![n-1]\!] = \{1\le l\le k\le n, l<n-1\}\cup\{1\le k<l\le n-1\} =: S_1\cup S_2,$$

to establish the result. From condition (b), for $(k,l)\in S_1$, $s_k\circ\sigma_l(\mathtt{Z}) = s_k(\mathtt{Z})$ and in particular $s_{n-1}\circ\sigma_l(\mathtt{Z}) = 0$ for $l\in[\![n-1]\!]$ while $s_n\circ\sigma_l(\mathtt{Z}) = 1$. From condition (a) and then condition (b), for $(k,l)\in S_2$, $s_k\circ\sigma_l(\mathtt{Z}) \le s_l\circ\sigma_l(\mathtt{Z}) = s_l(\mathtt{Z}) = 0$ and so $s_k\circ\sigma_l(\mathtt{Z}) = 0 = s_k(\mathtt{Z})$. □

Proof of Proposition 4. From Theorem 2 and the fact that $\nu^\psi = \nu$, for any $f, g \colon \mathsf{Z} \to [0,1]$

$$\begin{aligned}\int f(z)g(z')\nu(\mathrm{d}z)Q(z,\mathrm{d}z') &= \int f(z)g\circ\psi(z)\nu(\mathrm{d}z)\\ &= \int f\circ\psi^{-1}(z)g(z)\nu^\psi(\mathrm{d}z)\\ &= \int f\circ\psi^{-1}(z)g(z)\nu(\mathrm{d}z)\\ &= \int f(z')g(z)\nu(\mathrm{d}z)Q^*(z,\mathrm{d}z'),\end{aligned}$$

from which one can conclude. □

Proof of Lemma 5. Let $k \in \mathbb{Z}$, then the probability measure Λ^k has finite dimensional distributions satisfying, for $n \geq |k|$,

$$\Lambda^k_n(\mathrm{d}\mathsf{Z}) = \pi(\mathrm{d}z_k)\prod_{i=k+1}^{n} Q(z_{i-1},\mathrm{d}z_i)\prod_{i=-n}^{k-1} Q^*(z_{i+1},\mathrm{d}z_i),$$

which also guarantee the existence of Λ^k by Kolmogorov's Extension Theorem (Billingsley, 1995). Notice that for $n \geq |k|$ and $\mathsf{Z} \in S_k$, since (ν, Q, Q^*) is a reversible triplet,

$$\begin{aligned}&\nu(\mathrm{d}z_k)\prod_{i=k+1}^{n} Q(z_{i-1},\mathrm{d}z_i)\prod_{i=-n}^{k-1} Q^*(z_{i+1},\mathrm{d}z_i)\\ &\qquad = \nu(\mathrm{d}z_{k-\mathrm{sign}(k)})\prod_{i=k+1-\mathrm{sign}(k)}^{n} Q(z_{i-1},\mathrm{d}z_i)\prod_{i=-n}^{k-1-\mathrm{sgn}(k)} Q^*(z_{i+1},\mathrm{d}z_i),\end{aligned}$$

implying,

$$\begin{aligned}\Lambda^k_n(\mathrm{d}\mathsf{Z}) &= \varpi(z_k)\nu(\mathrm{d}z_k)\prod_{i=k+1}^{n} Q(z_{i-1},\mathrm{d}z_i)\prod_{i=-n}^{k-1} Q^*(z_{i+1},\mathrm{d}z_i)\\ &= \varpi(z_k)\nu(\mathrm{d}z_0)\prod_{i=1}^{n} Q(z_{i-1},\mathrm{d}z_i)\prod_{i=-n}^{-1} Q^*(z_{i+1},\mathrm{d}z_i)\\ &= \frac{\varpi(z_k)}{\varpi(z_0)}\pi(\mathrm{d}z_0)\prod_{i=1}^{n} Q(z_{i-1},\mathrm{d}z_i)\prod_{i=-n}^{-1} Q^*(z_{i+1},\mathrm{d}z_i)\\ &= \frac{\varpi(z_k)}{\varpi(z_0)}\Lambda^0_n(\mathrm{d}\mathsf{Z}),\end{aligned}$$

from which we conclude by application of Durrett (2019, Theorem 4.3.5), which is a mild generalization of Engelbert and Shiryaev (1980). □

Proof of Lemma 6. For $f, g \colon \mathsf{Z} \to [0,1]$ we have

$$\begin{aligned}\int f(z)g(z')\nu(\mathrm{d}z)\Psi(z,\mathrm{d}z') &= \int f(z)g\circ\psi(z)\nu(\mathrm{d}z)\\ &= \int f\circ\psi^{-1}(z)g(z)\nu^\psi(\mathrm{d}z)\\ &= \int f(z')g(z)\nu^\psi(\mathrm{d}z)\Psi^*(z,\mathrm{d}z').\end{aligned}$$

□

Proof of Lemma 7. Part (a) is clear from the definition of s_n. To establish parts (b) and (c) we use the decomposition

$$s_n(\mathsf{Z}, b) = g_{n-1}(z_{-\ell_n(b)}, \dots, z_{m_n(b)}) \vee f_{n-2}(z_{-\ell_n(b)}, \dots, z_{-\ell_n(b)+2^{n-2}-1}) \vee f_{n-2}(z_{-\ell_n(b)+2^{n-2}}, \dots, z_{m_n(b)}) \vee$$
$$g_{n-1}(z_{m_n(b)+1}, \dots, z_{r_n(b)}) \vee f_{n-2}(z_{m_n(b)+1}, \dots, z_{m_n(b)+2^{n-2}}) \vee f_{n-2}(z_{m_n(b)+2^{n-2}+1}, \dots, z_{r_n(b)}).$$

It follows that $s_n(\mathsf{Z}, b) \geq s_{n-1}(\mathsf{Z}, b)$, since if $b_n = 0$ then $[\![-\ell_n(b), m_n(b)]\!] = [\![-\ell_{n-1}(b), r_{n-1}(b)]\!]$ and so

$$s_{n-1}(\mathsf{Z}, b) = f_{n-2}(z_{-\ell_n(b)}, \dots, z_{-\ell_n(b)+2^{n-2}-1}) \vee f_{n-2}(z_{-\ell_n(b)+2^{n-2}}, \dots, z_{m_n(b)})$$

and if $b_n = 1$ then $[\![m_n(b)+1, r_n(b)]\!] = [\![-\ell_{n-1}(b), r_{n-1}(b)]\!]$ and so

$$s_{n-1}(\mathsf{Z}, b) = f_{n-2}(z_{m_n(b)+1}, \dots, z_{m_n(b)+2^{n-2}}) \vee f_{n-2}(z_{m_n(b)+2^{n-2}+1}, \dots, z_{r_n(b)}).$$

For the same reasons $s_n(\mathsf{Z}, b) \geq g_{n-1}(z_{-\ell_{n-1}(b)}, \dots, z_{r_{n-1}(b)})$. □

Proof of Lemma 8. The uniqueness of $b'_{1:n-1}$ follows from the fact that ℓ is uniquely determined by $b_{1:n-1}$ and $b'_{1:n-1}$ is uniquely determined by $\ell + k$. We have $\ell_{n-1}(b') = \ell + k$ and $r_{n-1}(b') = r - k$ by construction. Since for $i \in \mathbb{Z}$ $z'_i = z_{k+i}$, it follows that $z'_{-\ell_{n-1}(b')} = z_{k-\ell-k} = z_{-\ell}$ and $z'_{r_{n-1}(b')} = z_{k+r-k} = z_r$, so that indeed $[\![z'_{-(\ell+k)}, z'_{r-k}]\!] = [\![z_{-\ell}, z_r]\!]$. Since $b'_n = b_n$ and $\ell_{n-1}(b') = \ell_{n-1}(b) + k$, we also have $\ell_n(b') = \ell_{n-1}(b') + b_n 2^{n-1} = \ell_n(b) + k$ and similarly $r_n(b') = r_n(b) - k$, so $[\![z'_{-\ell_n(b')}, z'_{r_n(b')}]\!] = [\![z_{-\ell_n(b)}, z_{r_n(b)}]\!]$. Since $\tau(\mathsf{Z}, b) = n$, $s_n(\mathsf{Z}, b) = 1$ and $s_{n-1}(\mathsf{Z}, b) = 0$. Since $s_{n-1}(\mathsf{Z}, b)$ and $s_n(\mathsf{Z}, b)$ depend only on the values and the order of their inputs, and not the way they are indexed, we have $s_{n-1}(\mathsf{Z}, b) = s_{n-1}(\mathsf{Z}', b') = 0$ and $s_n(\mathsf{Z}, b) = s_n(\mathsf{Z}', b') = 1$. To conclude that $\tau(\mathsf{Z}', b') = n$, it remains only to show that $s_i(\mathsf{Z}', b') = 0$ for all $i \in [\![1, n-2]\!]$, but this is implied by Lemma 7-(b). □

Proof of Proposition 5. Let $\xi = (k, \mathsf{Z}^k) \in S_k \cap \{\mathsf{Z}^k \in \mathsf{Z}^k : \alpha_k(\mathsf{Z}^k) \wedge \alpha_k \circ \phi_k(\mathsf{Z}^k) > 0\}$. Then

$$\begin{aligned} r(\xi) &= \frac{\mathrm{d}\eta^{\phi_k}_{k,S_k}}{\mathrm{d}\eta_{k,S_k}}(\mathsf{Z}^k) \frac{\beta_k \circ \phi_k}{\beta_k}(\mathsf{Z}^k) \frac{\alpha_k \circ \phi_k}{\alpha_k}(\mathsf{Z}^k) \\ &= r_k(\mathsf{Z}^k) \frac{\alpha_k \circ \phi}{\alpha_k}(\mathsf{Z}^k) \\ &= 1, \end{aligned}$$

where we have used that $\alpha_k = r_k \cdot \alpha_k \circ \phi_k$ on S_k by part (b)i of Theorem 3, applied with $\mu = \eta_k \cdot \beta_k$ and $\phi = \phi_k$. □

B Measure theory tools

B.1 Standard results

Theorem 4 (Change of variables formula for Lebesgue measure)**.** *Let ϕ be a continuously differentiable, invertible function. If $f : \mathbb{R}^d \to \mathbb{R}$ is integrable then, with λ the Lebesgue measure,*

$$\int_{\mathbb{R}^d} f \circ \phi(\xi) \, |\mathrm{det}\phi'(\xi)| \, \lambda(\mathrm{d}\xi) = \int_{\mathbb{R}^d} f(\xi) \lambda(\mathrm{d}\xi),$$

where $\phi'(\xi)$ is the Jacobian matrix with entries $\phi'(\xi)_{ij} = \partial\phi_i/\partial\xi_j(\xi)$.

This is covered by Billingsley (1995, Theorem 17.2).

Example 31 (Jacobian of a linear mapping)**.** Consider the Lebesgue measure on $\big(\mathbb{R}, \mathscr{B}(\mathbb{R})\big)$ such that for any $a, b \in \mathbb{R}$, $a < b$ $\lambda\big((a,b]\big) = b - a$ and consider the scenario $\phi(\xi) = \alpha\xi$ where, without lost of generality, $\alpha > 0$. Recalling the definition $\lambda^{\phi^{-1}}(A) := \lambda\big(\phi(A)\big)$ for any $A \in \mathscr{B}(\mathbb{R})$ we have for $a, b \in \mathbb{R}, a < b$

$$\lambda^{\phi^{-1}}\big((a,b]\big) = \lambda\big((\alpha a, \alpha b]\big) = \alpha(b-a) = \alpha\lambda\big((a,b]\big),$$

that is $\lambda^{\phi^{-1}} = \alpha\lambda$ and $\lambda \equiv \lambda^{\phi^{-1}}$. We deduce on the one hand that

$$\begin{aligned}\int f(\xi)\lambda(\mathrm{d}\xi) &= \int f \circ \phi(\xi)\lambda^{\phi^{-1}}(\mathrm{d}\xi) \\ &= \int f \circ \phi(\xi)\alpha\lambda(\mathrm{d}\xi) \\ &= \int f \circ \phi(\xi)\mathrm{det}\phi'(\xi)\lambda(\mathrm{d}\xi)\end{aligned}$$

and using the Radon-Nikodym theorem (Theorem 1) we also have

$$\begin{aligned}\int f(\xi)\lambda(\mathrm{d}\xi) &= \int f \circ \phi(\xi)\lambda^{\phi^{-1}}(\mathrm{d}\xi) \\ &= \int f \circ \phi(\xi)\frac{\mathrm{d}\lambda^{\phi^{-1}}}{\mathrm{d}\lambda}(\xi)\lambda(\mathrm{d}\xi)\end{aligned}$$

and we deduce that, $\lambda-$almost everywhere,

$$\frac{\mathrm{d}\lambda^{\phi^{-1}}}{\mathrm{d}\lambda}(\xi) = \left|\mathrm{det}\phi'(\xi)\right|.$$

This result can be generalised to the multivariate scenario but also to nonlinear invertible and smooth mappings $\phi\colon \mathbb{R}^d \to \mathbb{R}^d$ by local linearisation.

B.2 Proofs

Proof of Proposition 2 . First observe that $\mathrm{d}\mu^\phi/\mathrm{d}\lambda^\phi = \rho \circ \phi$: for any $A \in \mathscr{E}$,

$$\begin{aligned}\int \mathbf{1}_A(\xi)\rho \circ \phi(\xi)\lambda^\phi(\mathrm{d}\xi) &= \int \mathbf{1}_A \circ \phi(\xi)\rho(\xi)\lambda(\mathrm{d}\xi) \\ &= \int \mathbf{1}_A \circ \phi(\xi)\mu(\mathrm{d}\xi) \\ &= \int \mathbf{1}_A(\xi)\mu^\phi(\mathrm{d}\xi).\end{aligned}$$

Then we find for $A \in \mathscr{E}$, $A \subseteq S$,

$$\begin{aligned}\int \mathbf{1}_A(\xi)\frac{\rho \circ \phi}{\rho}(\xi)\frac{\mathrm{d}\lambda^\phi}{\mathrm{d}\lambda}(\xi)\mu_S(\mathrm{d}\xi) &= \int \mathbf{1}_A(\xi)\frac{\rho \circ \phi}{\rho}(\xi)\frac{\mathrm{d}\lambda^\phi}{\mathrm{d}\lambda}(\xi)\mu_S(\mathrm{d}\xi) \\ &= \int \mathbf{1}_A(\xi)\frac{\rho \circ \phi}{\rho}(\xi)\frac{\mathrm{d}\lambda^\phi}{\mathrm{d}\lambda}(\xi)\rho(\xi)\lambda(\mathrm{d}\xi) \\ &= \int \mathbf{1}_A(\xi)\rho \circ \phi(\xi)\lambda^\phi(\mathrm{d}\xi) \\ &= \int \mathbf{1}_A(\xi)\mu_S^\phi(\mathrm{d}\xi),\end{aligned}$$

so that indeed $\mathrm{d}\mu_S^\phi/\mathrm{d}\mu_S = \frac{\rho\circ\phi}{\rho} \cdot \frac{\mathrm{d}\lambda^\phi}{\mathrm{d}\lambda}$. The proof that μ and μ^ϕ are mutually singular on $S^\complement$ follows the same arguments as in the proof of Theorem 3. □

C X-tra chance proof

This is a proof of the claims in Remark 27. Fix $u \in \mathbb{R}_+$, we show the result by induction. First we have $\beta_1(z) = 1 \times 1$, $\beta_1 \circ \phi_1(z) = 1$ and by considering $z \in S(\varpi_u, \varpi_u^{\phi_1})$ and $z \in S^{\complement}(\varpi_u, \varpi_u^{\phi_1})$ separately, and Theorem 6-6 we obtain $r_1(z) = \mathbb{I}\{u \leq \varpi\}\mathbb{I}\{u \leq \varpi \circ \phi_1(z)\} = \mathbb{I}\{u \leq \varpi(z) \wedge \varpi \circ \phi_1(z)\}$ and therefore with $a(r) = 1 \wedge r$ we deduce $\beta_2(z) = \beta_1(z)[1 - \mathbb{I}\{u \leq \varpi(z) \wedge \varpi \circ \phi_1(z)\}] = \mathbb{I}\{\varpi(z) \wedge \varpi \circ \phi_1(z) < u\}$. Assume that for some $k \in [\![2, n]\!]$ and any $z \in \mathsf{Z}$

$$\beta_k(z) = \mathbb{I}\{\varpi(z) \wedge \vee_{i=1}^{k-1} \varpi \circ \phi_i(z) < u\}.$$

From the assumption $\varpi \circ \phi_i \circ \phi_k = \varpi \circ \phi_{k-i}$ this implies

$$\beta_k \circ \phi_k(z) = \mathbb{I}\{\varpi \circ \phi_k(z) \wedge \vee_{i=1}^{k-1} \varpi \circ \phi_{k-i}(z) < u\} = \mathbb{I}\{\varpi \circ \phi_k(z) \wedge \vee_{i=1}^{k-1} \varpi \circ \phi_i(z) < u\}.$$

Therefore, proceeding as for $r_1(z)$ above and taking advantage of the fact that $\beta_k(\xi), \beta_k \circ \phi_k(z) \in \{0, 1\}$ we obtain

$$\begin{aligned}
r_k(z) &= \beta_k(z)\beta_k \circ \phi_k(z)\mathbb{I}\{u \leq \varpi(z) \wedge \varpi \circ \phi_k(z)\} \\
&= \mathbb{I}\{\varpi(z) \wedge \vee_{i=1}^{k-1} \varpi \circ \phi_i(z) < u \leq \varpi(z)\}\mathbb{I}\{\varpi \circ \phi_k(z) \wedge \vee_{i=1}^{k-1} \varpi \circ \phi_i(z) < u \leq \varpi \circ \phi_k(z)\}, \\
&= \mathbb{I}\{\vee_{i=1}^{k-1} \varpi \circ \phi_i(z) < u \leq \varpi(z)\}\mathbb{I}\{\vee_{i=1}^{k-1} \varpi \circ \phi_i(z) < u \leq \varpi \circ \phi_k(z)\} \\
&= \mathbb{I}\{\vee_{i=1}^{k-1} \varpi \circ \phi_i(z) < u \leq \varpi(z) \wedge \varpi \circ \phi_k(z)\}.
\end{aligned}$$

When $\vee_{i=1}^{k-1} \varpi \circ \phi_i(z) < \varpi(z) \wedge \varpi \circ \phi_k(z)$

$$1 - r_k(z) = \mathbb{I}\{u \leq \vee_{i=1}^{k-1} \varpi \circ \phi_i(z)\} + \mathbb{I}\{\varpi(z) \wedge \varpi \circ \phi_k(z) < u\}$$

therefore

$$\begin{aligned}
\beta_{k+1}(z) &= \mathbb{I}\{\varpi(z) \wedge \vee_{i=1}^{k-1} \varpi \circ \phi_i(z) < u\} \left[\mathbb{I}\{\varpi(z) \wedge \varpi \circ \phi_k(z) < u\} + \mathbb{I}\{u \leq \vee_{i=1}^{k-1} \varpi \circ \phi_i(z)\}\right] \\
&= \mathbb{I}\{\varpi(z) \wedge \vee_{i=1}^{k} \varpi \circ \phi_i(z) < u\} + \mathbb{I}\{\varpi(z) \wedge \vee_{i=1}^{k-1} \varpi \circ \phi_i(z) < u \leq \vee_{i=1}^{k-1} \varpi \circ \phi_i(z)\} \\
&= \mathbb{I}\{\varpi(z) \wedge \vee_{i=1}^{k} \varpi \circ \phi_i(z) < u\},
\end{aligned}$$

where we have used that $\mathbb{I}\{\varpi(z) \wedge \vee_{i=1}^{k-1} \varpi \circ \phi_i(z) < u\} = \mathbb{I}\{\vee_{i=1}^{k-1}\big(\varpi(z) \wedge \varpi \circ \phi_i(z)\big) < u\}$.

When $\vee_{i=1}^{k-1} \varpi \circ \phi_i(z) \geq \varpi(z) \wedge \varpi \circ \phi_k(z)$, using the same argument,

$$\begin{aligned}
\beta_{k+1}(\xi) &= \mathbb{I}\{\varpi(z) \wedge \vee_{i=1}^{k-1} \varpi \circ \phi_i(z) < u\} \\
&= \mathbb{I}\{\vee_{i=1}^{k}\big(\varpi(z) \wedge \varpi \circ \phi_i(z)\big) < u\} \\
&= \mathbb{I}\{\varpi(z) \wedge \vee_{i=1}^{k} \varpi \circ \phi_i(z) < u\},
\end{aligned}$$

which completes the proof.

D NUTS motivation

The criterion consists of stopping when $\|x_r - x_\ell\|^2$ reaches a stationary point, in the hope that it is a maximum. This requires the computation of a differential, that is the first order linear approximation of variations of $\|x_r - x_\ell\|^2$when x_ℓ (resp. x_r) is perturbed linearly $x_\ell + \epsilon v_\ell$ (resp. $x_r + \epsilon v_r$). This leads to

$$\|x_r - (x_\ell + \epsilon v_\ell)\|_2^2 - \|x_r - x_\ell\|_2^2 = -2\epsilon(x_r - x_\ell)^\top v_\ell + \epsilon^2 \|v_\ell\|_2^2,$$

and

$$\|x_r + \epsilon v_r - x_\ell\|_2^2 - \|x_r - x_\ell\|_2^2 = 2\epsilon(x_r - x_\ell)^\top v_r + \epsilon^2 \|v_r\|_2^2 ,$$

As $\epsilon \downarrow 0$ the dominant and linear term has coefficient

$$\lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \left\{ \|x_r - (x_\ell + \epsilon v_\ell)\|_2^2 - \|x_r - x_\ell\|_2^2 \right\} < 0 \iff (x_r - x_\ell)^\top v_\ell > 0,$$

and

$$\lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \left\{ \|x_r + \epsilon v_r - x_\ell\|_2^2 - \|x_r - x_\ell\|_2^2 \right\} < 0 \iff (x_r - x_\ell)^\top v_r < 0.$$

E Event chain algorithms

We briefly describe standard event chain processes for soft potentials and pairwise interactions. Define $x = (x_1, x_2, \ldots, x_m) \in \mathsf{X}^m$ with $\mathsf{X} = \mathbb{T}^d := -1/2 + \mathbb{R}^d/\mathbb{Z}^d$ and $v \in \mathsf{V} \subset \mathbb{R}^d$. The target distribution of interest has density

$$\gamma(x, v, i) = \gamma(x)\kappa(v) \propto \exp\big(-U(x)\big)\kappa(v)\mathbb{I}\{i \in [\![m]\!]\}$$

where γ has density with respect to the measure induced by the Lebesgue measure on $[-1/2, 1/2)^d$ Folland (1999, Chapter 6) and κ is the density with respect to the Lebesgue measure on $\mathsf{V} = \mathbb{R}^d$ or the Hausdorff measure on $\mathsf{V} = \mathbb{S}^{d-1}$. It is further assumed that $\kappa(-v) = \kappa(v)$ for $v \in \mathsf{V}$ and we focus on the scenario involving pairwise interactions,

$$U(x) := \sum_{1 \le i < j \le m} V(x_i - x_j),$$

where $V : (-1, 1)^d \to \mathbb{R}_+$ is continuously differentiable and such that $V(x_*) = V(-x_*)$ for all $x_* \in \mathsf{X}$. This leads to a probability density with exchangeability properties. The generator corresponding to event chain processes is given by

$$Lf(x, v, i) = \langle \nabla_x f, \mathbf{e}_i \circledast v \rangle + \lambda(x, v, i) \cdot \big[Rf(x, v, i) - f(x, v, i)\big] + \lambda_{\mathrm{ref}} \cdot \left[\int f(x, w, i)\kappa(\mathrm{d}w) - f(x, v, i) \right],$$

for $\lambda_{\mathrm{ref}} > 0$, $\{\mathbf{e}_i, i \in [\![m]\!]\}$ the canonical basis vectors and here $\circledast$ the Kronecker product. The intensity of the process is taken to be of the form

$$\lambda(x, v, i) = \sum_{j=1}^{m} \lambda_j(x, v, i)$$

with the convention $\lambda_i(x, v, i) = 0$ and for $j \neq i$, with $\langle \cdot, \cdot \rangle_+ := \max\{0, \langle \cdot, \cdot \rangle\}$,

$$\lambda_j\big(x, v, i\big) := \big\langle \nabla V_*(x_i - x_j), v \big\rangle_+,$$

and for $(x, v, i), (y, w, j) \in \mathsf{X} \times \mathsf{V} \times [\![m]\!]$,

$$R\big((x, v, i), \mathrm{d}(y, w, j)\big) := \sum_{k=1, k \neq i}^{m} \frac{\lambda_k(x, v, i)}{\lambda(x, v, i)} \delta_{(x, v, k)}\big(\mathrm{d}(y, w, j)\big).$$

This means that we follow trajectories of the form $t \mapsto (x_1, \ldots, x_{i-1}, x_i + tv, x_{i+1}, \ldots, x_m, v, i)$ with $t \ge 0$ for a random time arising from an inhomogeneous Poisson process of intensity $t \mapsto \lambda(x + t\,\mathbf{e}_i \otimes v, v, i) + \lambda_{\mathrm{ref}}$, a time at which one chooses between refreshing the velocity or selecting a new active particle randomly.

We check now that the corresponding process leaves the correct distribution invariant. We know that it is sufficient to show that $\mu(Lf) = 0$ for all functions $f : \mathsf{X} \times \mathsf{V} \times [\![m]\!] \to \mathbb{R}$ in a core of $\big(L, D(L)\big)$. Using Durmus, Guillin, and Monmarché (2018), it can be shown that the functions $f : \mathsf{X} \times \mathsf{V} \times [\![m]\!] \to \mathbb{R}$ such that for $i \in [\![m]\!]$, $f(\cdot, i) \in \mathsf{C}_b^2(\mathsf{X} \times \mathsf{V})$ (bounded support and twice continuously differentiable) define such a core. In fact with the isometric involution $\mathfrak{S}f(x, v, i) = f(x, -v, i)$, we can show the stronger property $\langle Lf, g\rangle_\mu = \langle f, \mathfrak{S}L\mathfrak{S}g\rangle_\mu$, for $f, g : \mathsf{X} \times \mathsf{V} \times [\![m]\!] \to \mathbb{R}$ such that the integral exists and where $\langle f, g\rangle_\mu := \int f g \mathrm{d}\mu$, which is the continuous time formulation of $(\mu, \mathfrak{S})-$reversibility Christophe Andrieu and Livingstone (2019). The property $\mu(Lf) = 0$ can be deduced by setting $g = \mathbf{1}$. We establish an intermediate result from which this latter property can be deduced.

Lemma 9. *Let $V : \mathsf{X} = \mathbb{T}^d \to \mathbb{R}_+$ be continuously differentiable and such that $V(x_*) = V(-x_*)$ for all $x_* \in \mathsf{X}$. Then for $i, j \in [\![m]\!]$, $i \neq j$*

(a) $\lambda(x, v, i) - \mathfrak{S}\lambda(x, v, i) = \big\langle \nabla_x U(x), \mathbf{e}_i \circledast v \big\rangle$,

(b) $\mathfrak{S}\lambda_j\big(x, v, i\big) = \lambda_i\big(x, v, j\big)$.

Proof. The first relation follows, for $i \in [\![m]\!]$, from

$$\begin{aligned}\sum_{j \neq i} \lambda_j\big(x, v, i\big) - \mathfrak{S}\lambda_j\big(x, v, i\big) &= \sum_{j \neq i} \big\langle \nabla_* V(x_i - x_j), v \big\rangle \\ &= \big\langle \sum_{j \neq i} \nabla_* V(x_i - x_j), v \big\rangle \\ &= \big\langle \nabla_x U(x), \mathbf{e}_i \circledast v \big\rangle.\end{aligned}$$

The second property follows from the assumption $V(x_*) = V(-x_*) = V \circ s(x_*)$ where $s(x_*) = -x_*$. Indeed, in this scenario the chain rule leads to $\nabla_* V(x_*) = (\nabla_* \circledast s)(\nabla V) \circ s(x_*) = -\nabla_* V(-x_*)$ and consequently for $i, j \in [\![1, m]\!]$, $i \neq j$

$$\begin{aligned}\mathfrak{S}\lambda_j\big(x, v, i\big) &= \big\langle -\nabla_* V(x_i - x_j), v \big\rangle_+ \\ &= \big\langle \nabla_* V(x_j - x_i), v \big\rangle_+ \\ &= \lambda_i\big(x, v, j\big).\end{aligned}$$

□

We now prove $\mu(Lf) = 0$. We can clearly ignore the refreshment component of the generator. An integration by part and Lemma 9 establish that

$$\begin{aligned}\int \langle \nabla_x f(x, v, i), \mathbf{e}_i \otimes v \rangle \mu\big(\mathrm{d}(x, v, i)\big) &= \int \langle \nabla_i f(x, v, i), v \rangle \mu\big(\mathrm{d}(x, v, i)\big) \\ &= \int f(x, v, i) \langle \nabla_i U(x), v \rangle \mu\big(\mathrm{d}(x, v, i)\big). \\ &= \int f(x, v, i) \left[\lambda(x, v, i) - \mathfrak{S}\lambda(x, v, i)\right] \mu\big(\mathrm{d}(x, v, i)\big) \\ &= \int f(x, v, i) \sum_{j \neq i} \left[\lambda_j\big(x, v, i\big) - \mathfrak{S}\lambda_j\big(x, v, i\big)\right] \mu\big(\mathrm{d}(x, v, i)\big) \\ &= \int \sum_{j \neq i} \lambda_j\big(x, v, i\big) \left[f(x, v, i) - f(x, v, j)\right] \mu\big(\mathrm{d}(x, v, i)\big).\end{aligned}$$

and we conclude by noting that

$$\int \lambda(x,v,i)\cdot\big[Rf(x,v,i)-f(x,v,i)\big]\mu\big(\mathrm{d}(x,v,i)\big)=\int\sum_{j\neq i}\lambda_j(x,v,i)\big[f(x,v,j)-f(x,v,i)\big]\mu\big(\mathrm{d}(x,v,i)\big).$$

The same calculations can be used for higher order interactions Harland et al. (2017).